Physiological Modulation of Learning and Decision-Making

Maaike M. H. van Swieten

Keble College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2020

Acknowledgements

Firstly, I would like to express my deep gratitude to Prof. Rafal Bogacz. His ever-positive attitude, deep insights, and invaluable feedback has not only helped cultivate this work but really encouraged my growth as a researcher and ultimately aided my understanding of computational techniques.

I am grateful for Prof. Peter Magill's guidance throughout the entirety of the DPhil; his advice has been indispensable. I am also very thankful to Prof. Sanjay Manohar for his suggestions and experimental support, which helped shape the direction of this work.

I would like to thank the Medical Research Council for their financial support, which enabled me to present my work at international conferences and attend summer schools to broaden my skill set.

Finally, I am very grateful to my close friends and family who have encouraged and supported me every step of the way.

Abstract

Many of the decisions we make in our day-to-day lives are influenced by internal physiological states and external environmental factors. In my thesis, I employed an interdisciplinary approach to investigate how internal physiological states, such as hunger, affect various aspects of decision-making. Neurobiological research has shown that food deprivation enhances dopaminergic signalling and increases the incentive salience of food. However, classical reinforcement learning theory only maps learning of external factors onto dopaminergic signalling. I extended the reinforcement learning theory in a biologically relevant manner to incorporate mechanisms by which internal physiological signals interact with the dopaminergic system to influence learning and action selection (chapter 3). This new theory predicts that food deprivation enhances learning, promotes approach behaviour, and reduces avoidance behaviour by enhancing the dopaminergic activation and teaching signals. I tested these theoretical predictions for decision-making in human volunteers. This study showed that hunger increased the valuation of food, but not of monetary rewards or non-food items (chapter 4). Hunger reduced the biases for approach and avoidance behaviour that sated people exhibit by modulating the trade-off between positive and negative consequences of actions (chapter 5). Hunger also interacted with economic choice, but only when outcomes of choices had to be learned, not when they were explicitly described (chapter 6). Lastly, this study showed that hunger enhanced the reliance of "gut-instinct", without affecting cognitive control (chapter 7). In summary, this thesis characterises mechanisms by which motivational drives, such as hunger, could affect valuation, learning, and action selection in a biological manner. It provides empirical evidence that hunger also affects decision-making for monetary rewards, which could have significant implications for both real-world economic transactions and for aberrant decision-making in eating disorders and obesity.

Contents

List of Figures **vi**

List of Tables **viii**

List of Abbreviations **ix**

1 Introduction **1**

2 Background and literature review **5**

2.1 Introduction 6

2.2 Decision-making 6

Learning 6

Valuation of outcomes 7

2.3 Neural correlates of decision making 12

Cortico-basal ganglia circuitry 12

Neural substrate of learning of subjective value 14

Value representation 16

2.4 Neural control of energy homeostasis 18

2.5 Theories of decision-making 22

Economic valuation 22

Reinforcement learning 24

Models grounded in the basal ganglia architecture 29

Action selection 31

Homeostatic control theories 32

2.6 Thesis aims 36

3 Modelling motivation in choice and learning 38

3.1 Introduction 39

Effects of motivation on dopaminergic responses 42

3.2 Results 44

Normative theory of state-dependent utility 44

Simulating state-dependent dopaminergic responses 47

Accounting for positive and negative consequences of actions . . 49

Neural implementation 51

Relationship to experimental data 61

Comparison to an existing incentive salience model 69

3.3 Discussion 72

Experimental predictions 72

Relationship to experimental data and implications 73

Relationship to other computational models 77

3.4 Methods 80

Simulating state-dependent dopaminergic responses 80

Simulating Go and Nogo learning 82

Inductive bias allows for faster learning 82

Influence of physiological state on action selection 83

State-dependent valuation 84

4 General experimental setup 85

4.1 Study design 86

Participants 86

Psychological questionnaires 86

Manipulation of metabolic state 87

Statistics 91

Software 91

Payment 91
4.2 Effects of fasting on hunger, valuation and attention 91
4.3 Computational modelling 92
Maximum likelihood estimation 93
Hierarchical models 93
Parameter recovery 96
Model comparison 96

5 Metabolic state and approach-avoidance behaviour 98
5.1 Introduction 99
5.2 Methods 102
Modified approach-avoidance learning paradigm 102
Behavioural Analyses 105
5.3 Results 105
Theoretical predictions for approach-avoidance learning 105
Food deprivation attenuated avoidance behaviour 111
Food deprivation did not attenuate learned values 114
5.4 Discussion 121
Comparison of computational models 122
Effect on learning 124
Effect on action selection 125
Neural substrates 126
Limitations 128
Conclusions 128

6 Metabolic state and risk-taking 130
6.1 Introduction 131
6.2 Methods 135
Experimental design 135
Behavioural analyses 138

Computational modelling 139
6.3 Results 139
Participants learned different reward distributions 139
Food deprivation altered learned risks, but not described risks . 140
Modelling of risk-sensitive choice behaviour 142
Computational modelling captured risk preferences 146
Reaction times did not explain changes in risk attitudes 149
6.4 Discussion 150
Experience versus description 151
Effects of food deprivation 152
Neural processing of risky decision-making 155
Conclusion 158

7 Metabolic state and learning strategies 160
7.1 Introduction 161
7.2 Methods 163
Paradigm 163
Stay-Switch Behaviour 164
Computational modelling 165
7.3 Results 167
Food deprivation only modulated model-free control 167
Food deprivation enhanced performance 170
Other reinforcement-driven behaviours were not affected 170
Computational modelling confirmed model-free effects 172
7.4 Discussion 176
Learning strategies 176
Neural substrates of model-free and model-based control 178
Limitations 180
Conclusion 181

8 General discussion **182**

8.1 Summary 183

8.2 Comparison of experimental and theoretical work 188

8.3 Considerations and future directions 191

Bibliography **194**

List of Figures

2.1 Parallel neural circuitries involved in decision-making 14
2.2 Modulation of value by risk 23
2.3 Computation in reinforcement learning 27

3.1 State-dependent modulation of dopaminergic responses 43
3.2 State-dependent computation of utility 45
3.3 Simulations with state-dependent reward prediction errors . . . 48
3.4 Schematic of utility computation in the basal ganglia network . 52
3.5 Go and Nogo weights for varying levels of motivation and reward size 60
3.6 Reward prediction error as a function of learning iteration . . . 62
3.7 Simulation of state-dependent dopaminergic responses using models grounded in the basal ganglia architecture 63
3.8 Simulation of the effects of salt-deprivation on behaviour 64
3.9 Simulation of state-dependent learning and choice 67
3.10 Learning of Go weights as a function of motivation 68
3.11 Comparison with incentive salience models 71

4.1 General study design 88
4.2 Hunger assessment 92

5.1 Task design approach-avoidance behaviour 103
5.2 Learned synaptic weights for Go and Nogo neurons 107
5.3 Predicted and empirical approach-avoidance behaviour at test . 110

5.4 Slowing and reaction times 112
5.5 Approach-avoidance behaviour divided by difficulty 113
5.6 Performance during training 115
5.7 Stay-switch behaviour after positive or negative feedback 116
5.8 Computational modelling 120
5.9 Learned Q-values 121

6.1 Task design risk-taking 136
6.2 Risk attitudes for learned and described risks 141
6.3 Model fitting risk-sensitive model 147
6.4 Computational processes captured risk preferences 148
6.5 Learned mean and spread of stimuli 149
6.6 Reaction times 150

7.1 Task design sequential decision-making 162
7.2 Stay-switch behaviour 168
7.3 Model-based and model-free control 169
7.4 Other reinforcement-driven actions 171
7.5 Computational modelling results 173
7.6 Schematic of value computation 174

List of Tables

5.1 Update rules computational models 118

7.1 Statistics stay behaviour 168
7.2 Parameters original model 175
7.3 Parameters re-parametrised model 176

8.1 Summary of experimental and theoretical findings 184
8.2 Comparison Theory and data 189

List of Abbreviations

ACC anterior cingulate cortex

ANOVA Analysis of Variance

Arc arcuate nucleus

BG basal ganglia

BIC Bayesian information criterion

BLA basolateral amygdala

CCK cholecystokin

CeM centromedial amygdala

CS conditioned stimulus

D1 dopamine type 1

D2 dopamine type 2

DA dopamine

DS dorsal striatum

EM expectation maximisation

EUT expected-utility theory

GLP-1 glucagon-like peptide 1

ITI inter-trial interval

LH lateral hypothalamus

MB model-based

MF model-free

MLE maximum likelihood estimation

mPFC medial prefrontal cortex

MSN medium spiny neurons

NAc nucleus accumbens

OFC orbitofrontal cortex

OpAL Opponent Actor Learning

PD Parkinson's disease

PT prospect theory

PYY peptide tyrosine tyrosine

RPE reward prediction error

RW Rescorla-Wagner

SEM standard error of the mean

SNc substantia nigra pars compacta

TD temporal difference

US unconditioned stimulus

VS ventral striatum

VTA ventral tegmental area

It's in making decisions that we learn to decide.
— Paulo Freire

1 Introduction

Hunger drives appetite and eating, while satiety halts them. However, sometimes we eat when we are not hungry or cease to eat before we are sated. Thus, internal physiological signals that indicate energetic needs *and* external environmental signals that may not be directly related to energy homeostasis control our food consumption. Environmental signals, such as reward size, sensory, cognitive, emotional, social, and experiential inputs play an increasingly more dominant role in the control of food intake. Impaired integration of physiological and environmental signals has been linked to the increasing prevalence of eating disorders and obesity. Conversely, internal physiological signals may also control the consumption of rewards that do not directly contribute to adequate energy homeostasis. These interactions could have significant implications for economic transactions.

One of the fundamental questions in value-based decision-making is how the brain orchestrates the integration of physiological and environmental signals to optimise behaviour that ensures survival and reproduction. For example, for long-term nutritional planning, individuals incorporate information about their metabolic need, the available resources in the environment, and the cost in time

and effort it takes them to obtain these resources. They learn to predict the availability of resources based on the context and estimate their value based on how useful they are in promoting survival. Finally, they opt for the choice that maximises their chance of survival. Although empirical and theoretical work have provided great insight, many questions still remain.

In my thesis, I employ an interdisciplinary approach to answer the following question: how do physiological states, such as hunger, affect various aspects of decision-making? To gain understanding, I first develop a biologically relevant computational model that provides a mechanistic account of how physiological signals may influence the evaluation and learning of outcomes as well as action selection. This model is different from previously developed models, because it brings together concepts from various fields, including economic utility theory, reinforcement learning, and models of motivation, with biology. It is grounded in empirical behavioural and neural data and is mapped onto the brain's architecture. This theoretical work provides new experimental predictions, which I address with an empirical study in human volunteers. In the experiments I designed, I mainly focus on the effects of food deprivation, or hunger, on valuation, learning and action selection. The study in human volunteers is valuable, because most empirical evidence on the effects of food deprivation on decision-making comes from animal studies. In humans, I can examine the effect of hunger on decision-making processes for food outcomes, but I can also explore how hunger influences decision-making for outcomes that are not directly beneficial to the current physiological state, such as obtaining money when hungry. This study highlights the generalising effects of hunger in complex decision-making processes.

The structure of my thesis is as follows. In chapter 2, I review existing literature on value-based decision-making. I discuss how external factors impact decision-making, and how the brain integrates metabolic signals from peripheral organs to generate adaptive behavioural responses to changing energy

demands. I review the learning theory, the underlying neural processes, and the computational models that have been developed to account for the effects.

In chapter 3, I provide a biologically relevant mechanistic account for the integration of internal and external factors in value-based decision-making. I provide theoretical insight for how the physiological state affects both learning and action selection processes. The proposed theoretical work is consistent with empirical data and the neural underpinnings of overeating and obesity. The theoretical work provides new predictions that can be tested experimentally. In chapter 4, I describe the general experimental setup and design that I developed to test how hunger influences various aspects of decision-making. I also provide methods for data analyses and computational modelling that are used in the following chapters.

Chapters 5–7 each cover one experimental study, investigating the influence of hunger on decision-making. All three chapters provide insight into the effect of hunger on learning and action selection, and the computational processes that underlie these effects. Chapter 5 addresses how hunger affects approach and avoidance behaviour. The trade-off between positive and negative consequences of choices is critical for survival. This study shows how food deprivation shifts the preference between approach and avoidance behaviour. Chapter 6 describes how hunger influences risk preferences. This study provides an unifying account for the effects of food deprivation on risk-taking when risks are presented in different ways. Chapter 7 reveals the effects of hunger on the contribution of reinforcement or cognition-guided action. Empirical evidence shows that hunger could influence either of these processes, but no study has provided insight into the trade-off between these two approaches when both can be employed simultaneously.

Finally, chapter 8 provides an overall discussion, in which I reflect on the theoretical predictions, and compare them with my empirical findings and existing literature.

Publications

Chapter 3 is adapted from van Swieten and Bogacz, 2020.

van Swieten, M.M.H., Bogacz, R. (2020) Modeling the effects of motivation on choice and learning in the basal ganglia. *PLoS Computational Biology* 16(5): e1007465.

A little knowledge is a dangerous thing. So is a lot.
— Albert Einstein

2

Background and literature review

Contents

2.1 Introduction 6
2.2 Decision-making 6
Learning 6
Valuation of outcomes 7
2.3 Neural correlates of decision making 12
Cortico-basal ganglia circuitry 12
Neural substrate of learning of subjective value 14
Value representation 16
2.4 Neural control of energy homeostasis 18
2.5 Theories of decision-making 22
Economic valuation 22
Reinforcement learning 24
Models grounded in the basal ganglia architecture 29
Action selection 31
Homeostatic control theories 32
2.6 Thesis aims 36

2.1 Introduction

In this chapter, I discuss the behavioural, neurobiological, and theoretical processes involved in decision-making. Decision-making is thought to encompass various processes, including the identification of internal and external states; the valuation of available actions; the selection of actions based on their valuation; the evaluation of outcomes after an action has been made, and learning in order to update the accuracy of these processes to improve future decision-making (Rangel et al., 2008). In the first section, I mainly focus on learning and the valuation of outcomes with respect to external and internal factors. Then, I discuss the neural correlates of learning and action selection, followed by the integration of metabolic signals and their impact on decision-making processes. Lastly, I discuss the theories that have been developed for economic decision-making, reinforcement learning and homeostatic control. I finish this chapter by addressing the aims of my thesis.

2.2 Decision-making

Learning

Humans adapt their behaviour to their external environment in a process often facilitated by learning. Human learning is a complex phenomenon that requires the acquisition of knowledge and the flexibility to modify existing neural processes to drive new patterns of desired behaviour. Fundamentally, learning is about associations between items and rewards, stimuli and responses, or actions and outcomes. Associative learning may occur via Pavlovian conditioning (classical conditioning), which depends on temporal contiguity, or instrumental (operant) conditioning, which depends on response-contingent reinforcement and temporal contiguity. Learning can occur consciously via explicit instructions or can occur unconsciously (implicitly) by experience as one perceives statistical regularities in our external world. Learning may occur with or with-

out feedback on the correctness or usefulness of one's actions, which is typically referred to as a reinforcement or reward.

In 1927, Palov was the first to describe the type of associative learning that we now know as Pavlovian conditioning. He presented dogs with a neutral cue (a tone) followed by the delivery of food. After repeated exposure to these two events, dogs salivated at the presentation of the cue, even in the absence of food. The once neutral stimulus had become predictive of future reward, and an association between a stimulus and an outcome was formed. Pavlovian conditioning is involuntary and does not require action selection. Conditioned Pavlovian responses comprise a small set of innate, or 'hard-wired', responses that are evolutionarily important. Pavlovian conditioning is an adaptive trait that presumably also occurs in natural environments: the ability to identify cues that predict future rewards and inform about upcoming threats are crucial for survival. Through conditioning, cues become incentive, promote actions and bias behaviour toward or away from particular outcomes (Berridge, 2012). Unfortunately, learned cues for rewards are also potent triggers of desires that promote maladaptive behaviour, such as overeating in the the presence of food (Watson et al., 2014).

Instrumental conditioning was first illustrated by Skinner, who designed an operant chamber, also known as a Skinner box, consisting of a lever and a food magazine (Skinner, 1959). Animals quickly learn that a lever press results in the delivery of a food pellet. The rewarding consequence of an action reinforces this action in the future, while a negative consequence prevents taking this action. Learning via instrumental conditioning strengthens the association between an action and its outcome. Typically, action-outcome contingencies are dependent on "discriminative stimuli", such as the current environment, or context.

Valuation of outcomes

Physiological needs (e.g. food and water) produce motivational drives (e.g. hunger and thirst) that engage goal-directed behaviours to promote survival. It is goal-

directed behaviour, because individuals adaptively change their behaviour in response to the same external stimuli when the internal state changes. In other words, the valuation of outcomes changes when the motivational drive changes.

Internal factors

Studies that include shifts in motivational drive provide insight into the motivational processes in action selection. For example, Mollenauer (1971) trained rats to run from a start box to a goal box to obtain a food reward. Animals in the deprived group were trained 22 hours after their daily meal, whereas animals in the sated group were trained only 1 hour after their daily meal. After both groups underwent 74 training trials, the deprived group ran consistently faster than the sated group. To investigate whether animals ran faster because they were more motivated, or because they learned that food has a higher incentive value when hungry, Mollenauer trained two additional groups of rats. These groups were either trained in a deprived state and tested in the sated state or vice versa. If animals require experience to learn how to act in a particular state, animals require a few trials to adjust and learn the new value of the state. If the motivation is modulated, a change in deprivation state leads to a direct impact on performance. The group that included a shift in motivation from training to test showed an immediate change in performance. Animals that shifted from deprived to sated ran slower, whereas animals that shifted from sated to deprived ran faster.

These findings suggest that prior experience with all possible deprivation levels may not be necessary. However, experimental studies do not provide conclusive results. Some studies argue that prior experience is necessary to adequately adapt performance (Balleine, 1992; Davidson et al., 1997; Lopez et al., 1992), while others show that this is not the case (Berridge and Schulkin, 1989; Cone et al., 2016). A classical example of dynamic shifts without prior experience is salt-seeking, or sodium appetite, which is the result of salt deprivation (Berridge and Schulkin, 1989; Cone et al., 2016; Krieckhaus, 1970). Animals

conditioned with a salty solution in a non-deprived state of salt, readily shift their preference when they are tested in a deprived state of salt, even when the solution does not contain any sodium. Without prior experience with sodium deficiency, sodium need directly impacts preference and performance.

Many of these behaviours are reinforcer-specific and state-dependent. Eating while hungry is more useful than eating while thirsty. Studies that examined this dissociation trained thirsty animals on a conditioned stimulus for saline or sucrose solution, before switching the motivational state to hunger and training them on a lever press for food. When tested, animals responded more for food when hungry and more for a solution when thirsty (Balleine, 1994). Similar observations have been made for nutrient-specific responding. A selective devaluation of nutrients reduces responding for this reinforcer (Papageorgiou et al., 2016). These data suggest that reinforcers act in a domain-relevant and state-specific manner. However, this interpretation is perhaps oversimplified; many resources comprise a combination of nutrients and in some cases may be relevant to both the hunger and thirst domain. For example, a sucrose solution contains calories as well as liquid and may therefore engage goal-directed actions when hungry or thirsty (Balleine, 1994; Dickinson and Balleine, 1990).

External factors

In addition to internal motivational drives, external factors also modulate goal-directed performance. In an analogous experiment to the one designed by Mollenauer (1971), Franchina and Brown (1971) showed how reward magnitude affected performance. Franchina and Brown trained four groups of food deprived rats to run to a goal-box, in which each group received a different number of food pellets. One group received 1 pellet and another group received 12 pellets throughout training and testing. Animals trained and tested with 12 pellets ran faster than the group rewarded with 1 pellet. The remaining two groups were either trained with 12 pellets and tested with 1, or trained

with 1 and tested with 12 pellets. Analogous to the shift from low to high motivation by Mollenauer (1971), animals that shifted from 1 pellet to 12 pellets increased their running speed. In contrast, animals that shifted from 12 to 1 pellet reduced their speed, analogous to animals that shifted from high to low motivation.

Together, these studies provide clear evidence that the current motivational drive and the incentive value of a reinforcer influence goal-directed behaviour. However, the experiments discussed until this point used reinforcement schedules that led to a reward on each trial. While this approach has provided invaluable insight, natural environments are rarely fully deterministic. These studies may provide a simplified interpretation of whether an outcome is good or bad, but provide limited insight into how relevant it is to take a certain action when choosing between alternatives. For example, resources may not always be available, or smaller immediate rewards may be preferred over delayed larger rewards in the face of starvation or imminent death. Importantly, how good or bad an outcomes is, is not absolute, but subjective. Outcomes are a function of various factors that I discuss in the following paragraphs. I provide a mathematical account for how valuation occurs and impacts decision-making in the theoretical section of this chapter.

Costs and benefits The benefit of an action must outweigh the cost of obtaining it, where the costs of an action may refer to the cost of time or the cost of effort (Day et al., 2010). Animals and humans typically prefer immediate or low effort rewards over delayed or high effort rewards of the same magnitude (Day et al., 2010; Walton et al., 2006). The ability to integrate the costs and benefits of potential courses of action is severely impaired in several common disorders, such as depression and schizophrenia. Empirical studies provide evidence that rewards are discounted hyperbolically when the time to obtain the reward increases. Each additional delay in receiving a reward has a large effect for short delays, but a small effect for long delays (Frederick et al.,

2002; Kable and Glimcher, 2007; Klein-Flügge et al., 2015). The discounting function for physical effort may be parabolic (Hartmann et al., 2013; Klein-Flügge et al., 2015).

Motivational drives influence effort-based decision making. Choices reflect not just the reward rate associated with the course of action but also metabolic costs of the actions themselves (Bautista et al., 2001). Animals trained on a two alternative choice task with options that differed in value and cost show a shift in preference resulting from a shift in food deprivation. Deprived animals are willing to exert a higher effort for a larger reward, whereas sated animals prefer low value-low cost alternatives (Kacelnik and Marsh, 2002; Mai et al., 2012; Marsh et al., 2004; Salamone et al., 1991). Motivational drives also affect time-based decision-making. Under deprivation, individuals show less self control, are more impulsive and more motivated, particularly for food rewards, but also for non-food related rewards (Bartholdy et al., 2016; de Ridder et al., 2014; Jewett et al., 1995; Kirk and Logue, 1997; Nederkoorn et al., 2009; Seibt et al., 2007; Skrynka and Vincent, 2019).

Uncertainty Many decisions involve uncertainty, which is defined as an imperfect knowledge about how choices lead to outcomes, not as the potential threat an action carries. Risk-taking is highly domain-specific: most people are not consistently risk-seeking or risk-averse across different decision-domains, such as financial, health, ethical, recreational and social decisions (Weber et al., 2002). The willingness to take risks is also subject to the wording of decisions, and whether risks involve gaining something positive or avoiding something negative (De Martino et al., 2006; Kahneman and Tversky, 1979). This is typically referred to as the decision context or the framing of choices. The subjective values are partially rescaled to the reward expected within a given context (Louie et al., 2015; Ludvig et al., 2014; Rigoli et al., 2016; Stewart, 2009)

When facing uncertain outcomes, animals and humans consider both the overall gain and the variability in available food sources and evaluate the options based on stored energy reserves (Bateson and Kacelnik, 1996; Brito e Abreu and Kacelnik, 1999; Caraco, 1981; Caraco and Lima, 1985; de Ridder et al., 2014; Kacelnik and Bateson, 1996; Levy et al., 2013; Shabat-Simon et al., 2018; Symmonds et al., 2010).

2.3 Neural correlates of decision making

The neural basis of action selection and reinforcement learning has largely focussed on the basal ganglia and its corresponding cortical areas. In the following sections, I discuss the basic functional anatomy of the cortico-basal ganglia circuitry and describe how decision-making processes may be mapped onto different brain areas.

Cortico-basal ganglia circuitry

The basal ganglia (BG) is a group of subcortical nuclei. The striatum is the main input nucleus of the BG and receives input from virtually all cortical regions, associated thalamic regions and the midbrain. Striatal neurons consist of a large group of GABAergic medium spiny neurons and small subsets of cholinergic and GABAergic interneurons. Medium spiny neurons can be grouped into at least two different subtypes based on the expression of dopamine type 1 (D1) or dopamine type 2 (D2) receptors and their projection targets (Gerfen and Surmeier, 2011; Smith et al., 1998). They either send an inhibitory input directly to the output nuclei of the BG, the globus pallidus internal segment and the substantia nigra pars reticulata, or indirectly, via the globus pallidus external segment and subthalamic nucleus. The direct or Go pathway is associated with the initiation of movements, while the indirect or Nogo pathway is associated with the inhibition of movements (Kravitz et al., 2010). The striatal Go neurons mainly express D1 receptors that are excited by dopamine, while

the striatal Nogo neurons mainly express D2 receptors that are inhibited by dopamine (Surmeier et al., 2007). Dopamine controls the competition between these two pathways during action selection and promotes action initiation over inhibition. The output nuclei of the BG send inhibitory projections via the thalamus back to the cortex, completing the cortico-basal ganglia loop.

Parallel loops

The cortico-basal ganglia circuit is crucial for emotional, cognitive, and sensorimotor functions in health and disease (Gunaydin and Kreitzer, 2016; Marchand, 2010; Yin and Knowlton, 2006). This circuit comprises three topographically preserved loops: sensorimotor, associative and limbic loop, that are well suited for parallel processing of information (Gremel and Lovinger, 2017; Middleton and Strick, 2000) (Fig. 2.1). The sensori-motor loop includes the somatosensory and motor cortices and dorsolateral striatum; the associative loop includes the prefrontal associative cortices, such as the dorsolateral and medial prefrontal, anterior cingulate, orbitofrontal and parietal cortices, and the dorsomedial striatum; and the limbic loop comprises the orbitofrontal cortex (OFC) and medial prefrontal cortex (mPFC), the amygdala, hippocampus, anterior cingulate cortex (ACC), and the ventral striatum (VS).

Dopamine neurons are found in the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc). The striatum receives the densest dopamine input, although many other brain areas, such as the cortex and amygdala, receive dopamine input as well. Dopamine input to the striatum is organised into three loops. The dorsolateral striatum is reciprocally connected with the lateral SNc. The dorsomedial striatum receives input from the medial SNc and projects to both the medial and lateral SNc. The VS receives input from the VTA and projects back to the VTA and SNc (Haber et al., 2000; Joel and Weiner, 2000). These dopaminergic sub-circuits contribute differently to learning and action selection, because of how they are connected with the cortico-basal ganglia loops.

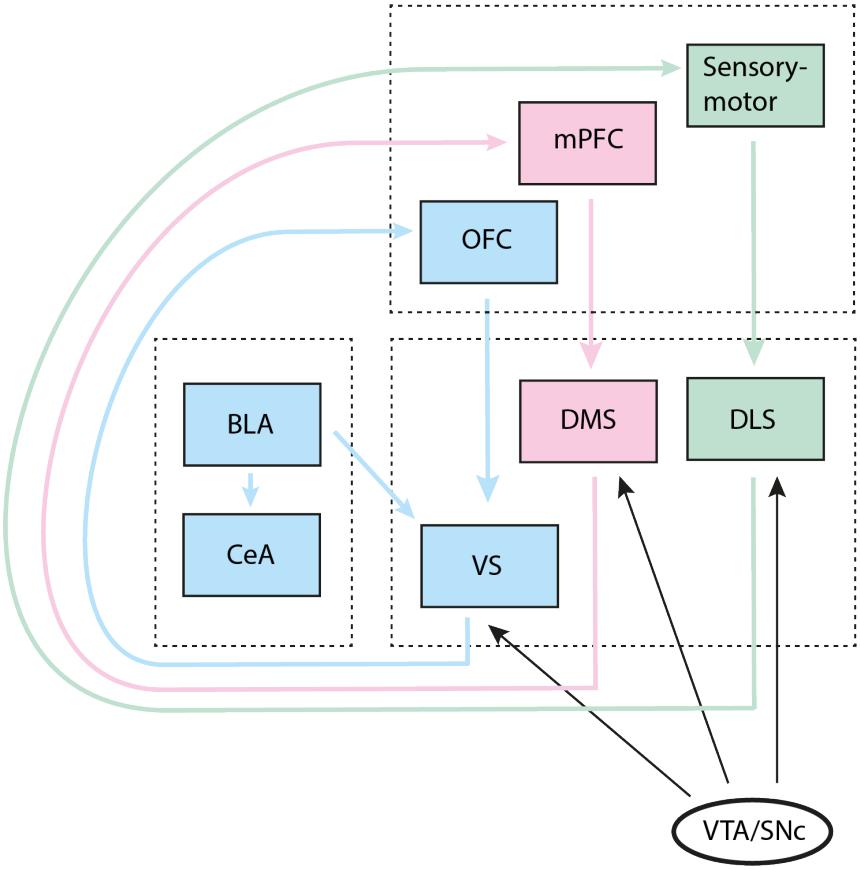


Figure 2.1: **Neural circuitries involved in decision-making**. The sensorimotor loop includes the sensory motor cortex and dorsal lateral striatum (DLS). The associative loop consists of the medial prefrontal cortex (mPFC) and dorsal medial striatum (DMS). The limbic loop links the orbitofrontal cortex (OFC) and ventral striatum (VS), with important contributions of the central (CeA) and basolateral nuclei of the amygdala (BLA). All loops receive dopaminergic inputs from the substantia nigra pars compacta (SNc) and ventral tegmental area (VTA).

Neural substrate of learning of subjective value

Experiential learning of the subjective value of actions is critical for optimal decision-making. The first compelling evidence for a neural substrate of learning was found by Schultz et al. (1997) in a Pavlovian conditioning paradigm. They recorded the activity of dopamine neurons while animals learned to associate a neutral cue with a reward delivery. They found that dopamine neurons respond to the receipt of unexpected rewards, but not the receipt of fully predicted rewards. After conditioning, dopamine neurons responded to the cue that reliably predicted upcoming rewards even though the neutral cue was not

inherently rewarding. Dopamine neurons showed a decrease in firing when a predicted reward was omitted. This study provides evidence that dopamine neurons in the midbrain encode a teaching signal that can be used to learn the subjective value of a reward. This teaching signal is also known as the reward prediction error, since it reflects the difference between the received reward and what was expected, and is a fundamental element in reinforcement learning models (discussed in section 2.5).

Subsequent studies demonstrated that dopamine responses scaled with the magnitude and the probability of a received reinforcement (Fiorillo et al., 2003; Tobler et al., 2005), but also with the delay and effort related costs associated with a reinforcement (Day et al., 2007; Fiorillo et al., 2008; Gan et al., 2009; Roesch et al., 2007) and reward proximity (Howe et al., 2013). Dopaminergic responses depend on the subjective utility of the obtained reward magnitude, rather than its objective magnitude (Stauffer et al., 2014) and are scaled by physiological states (Aitken et al., 2016; Cone et al., 2016; Papageorgiou et al., 2016). Pharmacological studies in healthy and Parkinson's disease (PD) patients demonstrated a causal role for dopamine in both learning and striatal BOLD prediction error signals. Participants who had received L-DOPA, a dopamine precursor, showed improved learning for selecting a more rewarding option, but not for avoiding a more punishing option (Frank et al., 2004; Pessiglione et al., 2006). Enhancement of dopamine resulted in larger striatal BOLD reward prediction error signals. Differences in this signal, compared to pharmacological blockade of dopamine receptors, could account for differences in the speed of learning when incorporated into reinforcement models (Pessiglione et al., 2006).

Together, dopamine seems to have two distinct functions in value-based decision-making. First, the level of dopamine controls the activation of the BG network by modulating the excitability of neurons (Hernández-López et al., 2000; Moyer et al., 2007; Thurley et al., 2008). Second, dopamine facilitates learning by triggering synaptic plasticity (Shen et al., 2008).

Value representation

Decision-makers often face options that have more than one dimension, such as climbing a tree to obtain a juicy apple. In order to make comparisons between options that are seemingly incomparable, the brain must integrate all dimensions into a single, abstract measure of its subjective value, and choose the option that is the most valuable in the current context. Subjective values may be divided into actions values and abstract values. Decision could be taken based on the relative comparison between the subjective value of options.

Striatal neurons represent the subjective value of actions. In a two alternative free choice task where animals choose between two actions of varying reward probabilities, recordings from dorsal striatal (DS) neurons show that there are three types of task related responses (Kawagoe et al., 1998; Kim et al., 2009; Lau and Glimcher, 2008; Samejima et al., 2005). Action value neurons respond *before* an action is chosen. They track the value of one of the actions, not both, independently of whether it is chosen. Chosen value neurons typically respond near the time of reward receipt and track the value of a chosen action. Choice neurons produce a categorical response when a particular action is taken. The striatum encodes the value of potential actions in an absolute manner, pointing towards a role for striatal neuron activity in the valuation process of decision-making. Although most of the evidence suggests that the DS is involved in the evaluation of action value, a recent study found evidence that the DS is also important for comparing options. Neuronal activity represented the difference between the values of two actions (Cai et al., 2011), suggesting that the DS may also be directly involved in the selection process.

In contrast to action values, choices may also be made based on the abstract values of "goods". Cortical regions represent subjective values of different rewards. Evidence for absolute representation of reward value, independent of movement signals, has been found in the OFC (Padoa-Schioppa and Assad,

2006). In a task, monkeys choose between two types of juice offered in different amounts. OFC neurons show three types of responses, analogous to the ones found in the striatum. Offer value neurons increase their firing rate linearly with the subjective value of one of the rewards on offer. The subjective value of an offer is calculated based on the type and quantity of the juice. Chosen value neurons track the subjective value of the chosen option in a non-specific manner, independent of juice type and amount. Taste neurons produce a categorical response when a particular juice is chosen. All responses are independent of the motor response produced by the animal to make its choice and the spatial arrangement of the stimuli. Offer value and chosen value neurons respond after the presentation of option and at the time of the reward receipt. Taste neurons respond only at the time of reward receipt.

In addition to value representation, dorsomedial prefrontal and parietal cortical areas are also implicated in the comparison of values and the selection process (Platt and Huettel, 2008; Rushworth and Behrens, 2008). Relative value signals have been found in the lateral intraparietal cortex. Neurons increase their firing to actions that are more valuable, relative to other possible actions, increasing the chance of being selected for motor execution (Dorris and Glimcher, 2004; Platt and Glimcher, 1999; Sugrue et al., 2004). Action-based subjective values have also been found in the ACC (Rudebeck et al., 2008) and the ACC and mPFC encode the difference between action values (Wunderlich et al., 2009). Decision-related signals in these areas occur after the choice has been made, suggesting that these regions could play an important role in adjusting choice behaviour.

The insular cortex and amygdala play an important role in emotional decision-making and gating of homeostatic circuits (Calhoon et al., 2018; Livneh et al., 2017, 2020). The basolateral amygdala (BLA) encodes and the insular cortex retrieves outcome values to guide choice between goal-directed actions (Parkes and Balleine, 2013). The insular cortex encodes risk prediction errors and fram-

ing of choices (Clark et al., 2008; De Martino et al., 2006; Paulus et al., 2003; Preuschoff et al., 2008; Wright et al., 2013).

2.4 Neural control of energy homeostasis

The central nervous system integrates numerous metabolic signals from the periphery and thereby generates adaptive behavioural responses to changing energy demands. The orexigenic (appetite-stimulating) hormone ghrelin is released from the stomach, and the anorexigenic (appetite-suppressing) hormones insulin and leptin are released from the pancreas and fat cells, respectively. Other anorexigenic hormones released from the gut include, cholecystokin (CCK) and peptide tyrosine tyrosine (PYY) which is co-secreted with glucagon-like peptide 1 (GLP-1).

Classical studies from the 1950s have identified the hypothalamus as the key brain area in controlling food intake and regulating energy balance (Anand and Brobeck, 1951; Hoebel and Teitelbaum, 1962; Olds and Milner, 1954). It is positioned near a 'leaky' part of the blood-brain-barrier and therefore critical for the integration of hormonal and nutritional metabolic signals. The hypothalamus consists of various nuclei. The Arcuate nucleus (Arc) comprises two distinct and functionally antagonistic types of neurons. The orexigenic Neuropeptide Y and agouti-related peptide-expressing neurons drive feeding (Aponte et al., 2011; Krashes et al., 2011). They transmit negative valence, promote food seeking and reduce their activity once food is identified (Betley et al., 2015). The anorexigenic pro-opiomelanocortin-expressing neurons inhibit feeding (Vong et al., 2011).

The lateral hypothalamus (LH) plays an important role in the regulation of feeding behaviour and energy expenditure (Stuber and Wise, 2016). Electric lesion of the LH decreases feeding, drinking and body weight (Anand and Brobeck, 1951), likely by manipulating the rewarding properties of food (Olds and Milner, 1954). The LH receives information about peripheral energy re-

serves from hormones like leptin (Elmquist et al., 1998; van Swieten et al., 2014), nutrient signals such as glucose and afferents originating from energy sensing mechanisms in the Arc (Elias et al., 1998).

Sub-populations in the LH control food intake and body weight homeostasis via a number of sub-circuits. The glutamatergic projection to the lateral habenula (Stamatakis et al., 2016) reduces food intake and negative aspects of reward. The GABAergic and melanin concentrating hormone neurons enhance the valuation of food items and increase learning of nutritionally valuable food cues to drive food consumption (Jennings et al., 2015). GABAergic and orexin neurons enhance the dopaminergic input from VTA neurons projecting onto the nucleus accumbens (NAc) (Sheng et al., 2014). LH neurons projecting to the extended amygdala are involved in the motivation to obtain food intake (Jennings et al., 2013). The LH neurons projecting to the OFC and secondary taste area signal satiety (Simon et al., 2006) and control the hedonic and goal-directed values of food.

Hindbrain regions are reciprocally connected with the hypothalamic nuclei to control appetite (D'Agostino et al., 2016; Wang et al., 2015) particularly by controlling meal-size. Satiety signals such as leptin, CCK, GLP-1 and PYY induce satiety by dampening the activity in the reward system, reducing the value and palatability of food, to decrease food consumption (Alhadeff et al., 2012; Gong et al., 2020; Steinert et al., 2017; Volkow et al., 2011).

The hypothalamus and hindbrain are tightly connected with the cortico-limbic system, consisting of various cortical areas, the basal ganglia, hippocampus and amygdala. It provides emotional, cognitive and executive support for ingestive behaviour (Kelley et al., 2005). The neurobiology of reward, economics and decision-making play an important role in food intake (Murray and Rudebeck, 2013; Rangel, 2013)

Chronic food-restriction increases dopaminergic signalling (Carr et al., 2003; Stuber et al., 2002; Verhagen et al., 2009) by acting on ghrelin receptors in the

VTA and SNc (Zigman et al., 2006). Similarly, ghrelin administration also increases the level of dopamine in the midbrain (Andrews et al., 2009), and dopamine turnover in the DS and NAc (Abizaid et al., 2006; Anderberg et al., 2016; Cone et al., 2014; Kawahara et al., 2009). Ghrelin increases the motivation to obtain food and promotes feeding by enhancing the mesolimbic pathway to the NAc (Aberman et al., 1998; Abizaid et al., 2006; Cone et al., 2014; Han et al., 2018; Kawahara et al., 2009; Malik et al., 2008; Naleid et al., 2005). In contrast, devaluation paradigms by satiety show that dopamine responses for goal-directed and Pavlovian behaviour are reduced instantly (Aitken et al., 2016; Cone et al., 2016; Papageorgiou et al., 2016). This is consistent with the fact that leptin administration reduces dopamine release from the midbrain (Krügel et al., 2003; van der Plasse et al., 2015) via its receptors in the VTA (Elmquist et al., 1998; Figlewicz et al., 2003). Leptin regulates the reward value of nutrients (Domingos et al., 2011) and reduces food intake (Fulton et al., 2006; Hommel et al., 2006), presumably via a D2-receptor dependent mechanism (Billes et al., 2012; Pfaffly et al., 2010; Thanos et al., 2008). Obese individuals are leptin insensitive, have reduced levels of striatal D2-receptors (Hamdi et al., 1992; Pfaffly et al., 2010; Stice et al., 2008; Volkow et al., 2008; Wang et al., 2001, 2009), and striatal dopamine transporters (Chen et al., 2008).

Insulin also interacts with the dopaminergic system (Labouèbe et al., 2013; Mebel et al., 2012; Stouffer et al., 2015). Behaviourally, insulin alters anticipatory behaviour for food (Labouèbe et al., 2013), the hedonic value of food (Figlewicz et al., 2004; Mebel et al., 2012; Tiedemann et al., 2017), motivation for food rewards (Figlewicz et al., 2006), but not learning of food value (Figlewicz et al., 2004).

Separate BG circuitries are thought to mediate the hedonic and nutritional values of food. Following metabolic cues, D1-expressing neurons in the DS drive goal-directed action guided by the nutritional value of food. Dopamine

activity in the VS is linked to the hedonic value of food (Tellez et al., 2016).

Cortical areas integrate sensory inputs in a state-dependent manner. Hunger and the caloric content of food items modulate the activity in the posterior cingulate cortex, medial OFC, insula, and fusiform gyrus (Siep et al., 2009; Verhagen, 2007). However, sensory input, such as images of appetising food items, can also trigger the desire to eat, even in the absence of hunger (Spence et al., 2016). A food-deprived state enhances goal-directed behaviour by enhancing the dopamine utilisation in the mPFC (Carlson et al., 1987; Moscarello et al., 2007; Siep et al., 2009). Activity in the OFC is correlated to the subjective valuation of food items in a state-dependent manner (O'Doherty et al., 2000; Plassmann et al., 2007)

The amygdala is critical for interoception of homeostasis and the regulation of food intake (Critchley et al., 2004). Food deprivation increases c-Fos immunoreactivity in the BLA and centromedial amygdala (CeM) (Moscarello et al., 2009). BLA neurons projecting to the NAc preferentially encode positive valence, whereas BLA neurons projecting to the CeM preferentially encode negative valence (Namburi et al., 2015). The BLA plays an important role in satiety-dependent outcome devaluation (Corbit et al., 2013; Parkes and Balleine, 2013), and mediates hunger-induced invigorated responding (Malvaez et al., 2015). Satiety enhances the processing of negative valence and diminishes processing of positive valence, showing that homeostatic need alters the reward value by changing the balance between the two BLA circuitries (Calhoon et al., 2018). These data suggest that the BLA may play an important role in incorporating information about the homeostatic need to dynamically evaluate rewards and translate value into motivated behaviour.

2.5 Theories of decision-making

In this section, I discuss the theoretical frameworks developed for the understanding of various aspects of decision-making. I first discuss the economic utility theory that describes how uncertainty in reward outcomes impacts the subjective value attributed to a decision or an action. Then, I introduce the reinforcement learning framework and lastly, I discuss the theories developed for the integration of both internal and external factors to control homeostasis.

Economic valuation

Volatile environments are unpredictable. To choose the best course of action in such an environment, the subjective value of an action needs to reflect the uncertainty in outcomes.

The valuation of uncertain outcomes may occur via two ways. The brain either assigns value by computing its statistical moments, such as the mean, variance, skewness of a distribution (Preuschoff and Bossaerts, 2007) or the brain computes a utility value for each possible outcome, which is weighted by a function of the probabilities (Kahneman and Tversky, 1979; von Neumann and Morgenstern, 1944). The latter relies on economic theories such as expected-utility theory (EUT) or prospect theory (PT). Within economics, the term utility is used to model the total satisfaction derived from consuming a good or service (Friedman and Savage, 1952).

EUT is a normative model of decision-making in which the most optimal, or rational, decision should be made in a given situation. A fully rational decision-maker is motivated to maximise the currency of interest. In EUT, the value of a prospect equals $\sum_x p(x)v(x)$, where the $v(x)$ is the value of an outcome x, which is weighted by its objective probability, $p(x)$. The value function, $v(\cdot)$, is a concave function of outcomes, and the probability function is the identity function. Note that a special case that is often used in the

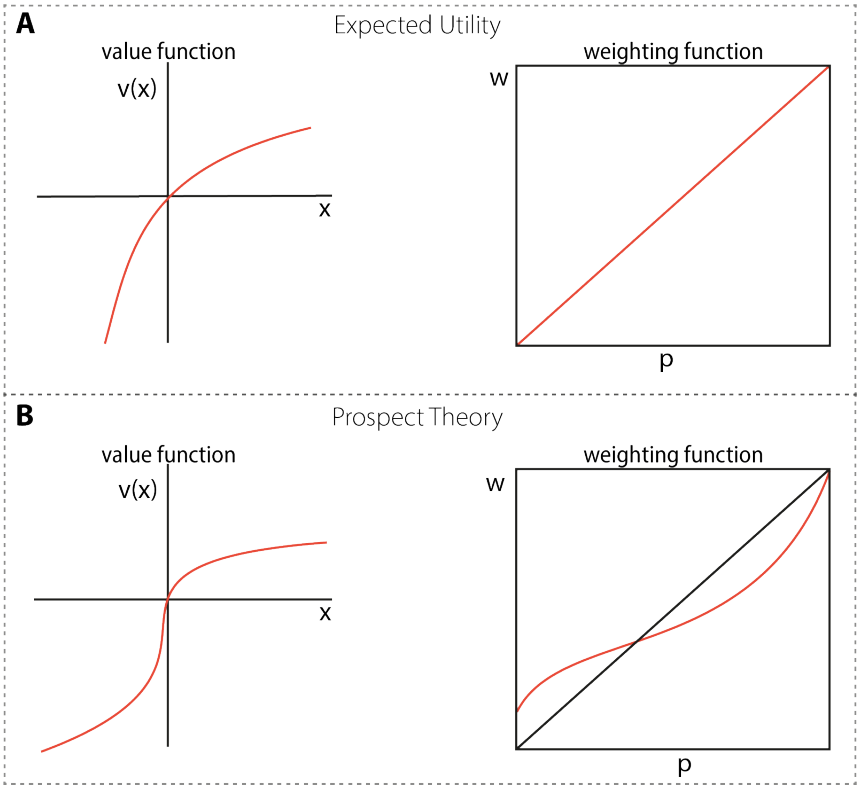


Figure 2.2: **Values are modulated by risk**. A) In expected-utility theory, the value of a prospect equals $\sum_x p(x)v(x)$, where the $v(x)$ is the value of an outcome x, which is weighted by its objective probability, $p(x)$. B) In prospect theory, the value of a prospect equals $\sum_x w(p(x))v(x - rp)$, where the values of the outcomes now depend on a reference point, rp, and are weighted by a nonlinear function, $w(\cdot)$, of the objective probabilities.

experimental neuroeconomics literature is $v(x) = x$, reducing the EUT function to the expected value of the prospect (Fig. 2.2A).

Empirical evidence suggests that humans may not be rational in their choice behaviour. PT provides insight into how people make 'irrational' decisions under uncertainty. In PT, the value of a prospect equals $\sum_x w(p(x))v(x - rp)$, where the outcome values depend on a reference point, rp, and are weighted by a nonlinear function, $w(\cdot)$, of the objective probabilities. Comparing prospects to a reference point can create framing effects. A different value is assigned to the same prospect depending on how it is positioned with respect to a cognitive reference point, creating different decision contexts. The value function is usually concave for gains (positive decision contexts), but convex for losses

(negative decision contexts) (Fig. 2.2B). PT also accounts for the tendency to prefer avoiding losses compared to acquiring equivalent gains: $v(x) < v(x)$ for $x > 0$. This property is better known as loss aversion, which is captured by a kink in the value function. In PT, outcomes with small probabilities are overweighted and outcomes with large probabilities are underweighted. Some of the behavioural data that is inconsistent with EUT, can be successfully explained by PT (Kahneman and Tversky, 1979).

Both EUT and PT models describe prospects of which all probabilities are known and described to participants. This is in contrast to situations where probabilities are not known and have to be learned (this is sometimes referred to as ambiguity rather than risk (Camerer and Weber, 1992)). Recent studies have shown that humans apply a different weighting function to outcomes that are learned through sampling compared to described outcomes. Outcomes below the reference point are underweighted, while outcomes above the reference point are overweighted (Hertwig et al., 2004; Ludvig et al., 2014; Ludvig and Spetch, 2011; Madan et al., 2014). Given the different probability weighting functions, learning under uncertainty may be different from valuation of uncertain prospects. This possibility is not always accounted for in theories on associative learning.

Reinforcement learning

In order to learn how to make good decisions, the brain needs to compute a value signal that measures the desirability of an outcome that was generated by its previous decisions. Models of reinforcement learning compute a prediction error signal after observing an outcome generated by an action or stimulus. This signal measures the quality of the forecast that was implicit in the previous valuation. Each time a learning event occurs, the value of an action or stimulus is changed by an amount proportional to the prediction error. After sufficient time, an individual learns to assign the correct value to an action or stimulus.

Dopaminergic signals in the brain are thought to encode the prediction error signal in both Pavlovian and instrumental conditioning.

In this section, I first discuss the models for Pavlovian conditioning followed by models for instrumental conditioning (Fig. 2.3), and I finish by discussing models of learning and action selection that are grounded in the basal ganglia architecture.

Pavlovian conditioning

Pavlovian conditioning involves learning, but not action selection. When an agent experiences a reward $r(s_t)$ associated with a stimulus s on a particular trial t that differs from what it expected based of previous learning $V(s_t)$, a reward prediction error arises (Bush and Mosteller, 1951):

$$\delta = r(s_t) - V(s_t). \tag{2.1}$$

Using this reward prediction error, the expectation on the next trial is updated according to:

$$V(s_{t+1}) = V(s_t) + \alpha\delta, \tag{2.2}$$

where $0 < \alpha \leq 1$ is the learning rate that weights the importance of recent outcomes. This model only updates the value of the presented stimulus $V(s_t)$, while the values of unobserved stimuli remain the same. Rescorla and Wagner extended this model to account for learning when multiple stimuli are presented simultaneously (Rescorla and Wagner, 1972). The state value is computed as the sum of all available stimuli i on trial t: $V(s_t) = \sum_i V(s_{t,i})$ and used to update observed stimuli using Eq. (2.2).

In the Rescorla-Wagner (RW) model an agent learns the association between a conditioned stimulus (CS) and an unconditioned stimulus (US), without considering the temporal relationship between them. This shortcoming was resolved in the temporal difference (TD) learning model (Sutton, 1988). In this algorithm, the agent aims to learn an estimate of a value function,

$V(s_t)$ for each arbitrary time step t. The function predicts all future rewards r, discounted by time within a trial, following the presentation of a stimulus s_t.

$$V(s_t) = E[r(s_t) + \gamma r(s_{t+1}) + \gamma^2 r(s_{t+2}) + \gamma^3 r(s_{t+3}) + \dots], \tag{2.3}$$

The expectation is over the uncertainty in reward delivery and state transitions. The preference for immediate over delayed rewards is captured by the temporal discounting factor $0 < \gamma \leq 1$. Instead of waiting until all outcomes are experienced, the TD model estimates future reward by repeating the following algorithm for each time step:

$$V(s_t) \leftarrow V(s_t) + \alpha[r(s_t) + \gamma V(s_{t+1}) - V(s_t)], \tag{2.4}$$

where $[r(s_t) + \gamma V(s_{t+1}) - V(s_t)]$ is the prediction or temporal difference error, and $\gamma V(s_{t+1})$ represents an estimate of upcoming rewards. The TD model is one of the leading models in conditioning. The reward prediction error that drives learning in this model corresponds with the phasic firing of dopaminergic neurons (Schultz et al., 1997).

An important difference between the RW and TD model is that the TD model provides a real-time account of learning, while the RW model provides a trial level account. The prediction error aligns with the presentation of the US. During conditioning this signal transfers to the CS. At the end of training, the CS predicts the expected value of upcoming rewards and the US response has disappeared, because the difference predicted rewards and the true outcome is zero.

Instrumental conditioning

In classical conditioning, the expected value of a state $V(s_t)$ is defined as the expected cumulative reward of one or more stimuli presented. The association between a CS and an US does not require the selection of an action. However, when an individual requires to choose an action, the value of an action in a state

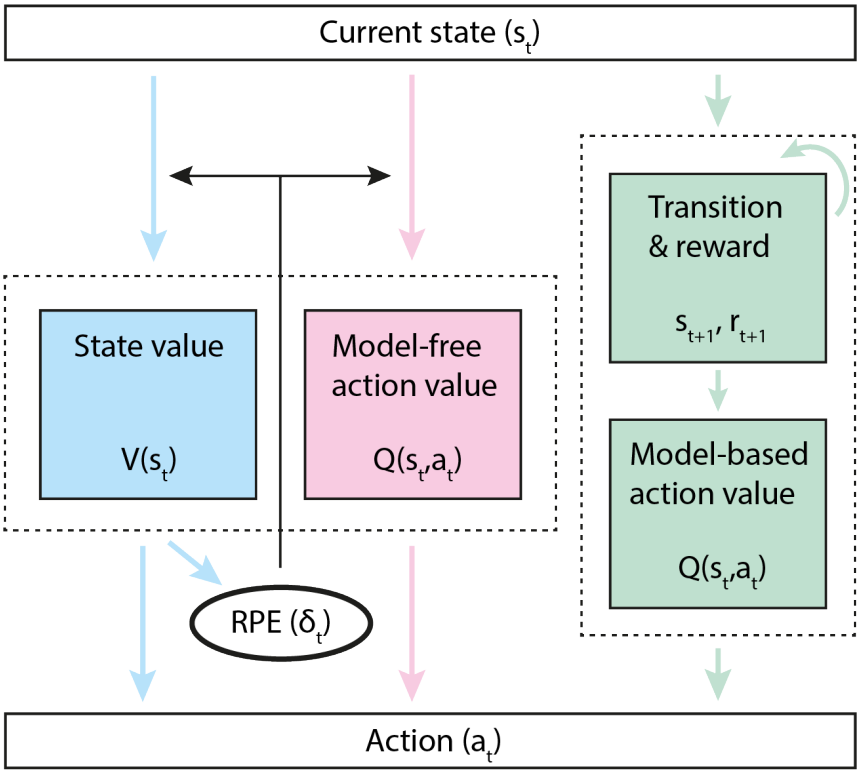


Figure 2.3: **Computation in reinforcement learning**. Classical condition involves learning without decisions. The reward prediction error (RPE) depends on the current reward, experienced on a trial, compared to the prediction associated with a stimulus $V(s_t)$, which is based on previous experience.

$Q(s, a)$ needs to be learned instead. An agent learns to assign values to actions following the structure of Markovian decision problems, which is defined by the transition function and reward function.

At each time step t, the model world takes on a discrete state s_t and the agent chooses an action a_t. Together, these probabilistically determine the next state, s_{t+1}. This transition function can be formally defined by:

$$T(s, a, s') = P(s_{t+1} = s' | s_t = s, a_t = a), \tag{2.5}$$

which specifies the likelihood that an action a in state s leads to a specific new state s'. Together with the reward function of the state that predicts all future, discounted rewards, the action value is estimated according to:

$$Q(s, a) = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \ldots | s_t, a_t], \tag{2.6}$$

where r_{t+k} denotes the reward that is received at time $t + k$. The notation $E[\cdot|s_t, a_t]$ reflects the expectation over all possible future sequences of states and rewards given an initial state and action.

The objective is to guide choice by predicting future rewards. All future rewards are captured in the value of the next state. Therefore, this function can be rewritten as follows:

$$Q(s, a) = R(s) + \sum_{s'} T(s, a, s') \max_{a'} Q(s', a') \tag{2.7}$$

The state-action value depends on the reward function R and the average over future states according to the transition function T. The agent should take the best action in the next state with value $\max_{a'}[Q(s', a')]$.

Values derived this way are typically referred to as model-based (Daw et al., 2005). The transition function T captures how particular actions in particular states lead to other states. The reward function R represents the rewarding value of outcomes. If the reward value of outcomes change (represented by the reward function R), Q-values and choices derived from this knowledge, will immediately reflect this change. This makes decisions sensitive to changes in response-outcome contingencies and value.

Instead of estimating future value, an estimate can also be obtained by averaging samples, which is done with the Q-learning algorithm (Watkins, 1989). An agent starts with an initial estimate of the value $Q(s_t, a_t)$. For each time step t in state s_t the agent receives reward r_t and takes action a_t. This action leads to a new state s_{t+1} according to a transition distribution $T(s_t, a_t)$ unknown to the agent. Using this experience the agent updates the estimate of the future value according to:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\delta_t, \tag{2.8}$$

where δ_t is the reward prediction error that measures the difference between the old estimate and the new estimate in a temporal manner:

$$\delta_t = r_t + \gamma \max_{a'}[Q(s_{t+1}, a')] - Q(s_t, a_t), \tag{2.9}$$

where α is the learning rate, γ is the discount factor and $\max_{a'}[Q(s_{t+1}, a')]$ is the estimate of the optimal future value.

The Q-learning algorithm learns action values that generate the optimal behaviour and converge to the objective expected values. At each step of the learning process the action with the largest value is selected. There are two other variants of this model: SARSA and actor-critic model. They rely on the same principles, but differ from Q-learning in their exact specification on the prediction error, and learning rules and are not guaranteed to converge to the optimal policy. All these models assume that the agent is ignorant of the transition or reward function, they are computationally simple and do not require the agent to keep track of long reward sequences to learn action values. These learning algorithms are typically seen as model-free algorithms.

Models grounded in the basal ganglia architecture

Recent developments in reinforcement learning models have led to the development of models grounded in the architecture of the basal ganglia (Collins and Frank, 2014; Mikhael and Bogacz, 2016; Möller and Bogacz, 2019). These models suggest mechanisms by which dopamine influences learning and action selection via the Go and Nogo pathway of the BG. Dopamine enhancement promotes approach behaviour, whereas dopamine reduction promotes avoidance behaviour. These models assume that Go and Nogo neurons encode positive consequences and negative consequences of actions, respectively.

Opponent Actor Learning model The Opponent Actor Learning (OpAL) model (Collins and Frank, 2014) includes two components: an actor that learns tendencies to select particular actions and a critic that learns an overall value

of the current context or state. Biologically, the critic represents computations in the VS and the actor represents computations in the DS.

The critic learns the value $V(s_t)$ of a state using the standard Rescorla–Wagner rule in Eq. (2.2). Instead of learning one state-action value $Q(s_t, a_t)$, the tendencies to choose or inhibit actions are encoded in the strengths or the synaptic connections between the cortical neurons associated with that state and the striatal Go (G) and Nogo (N) neurons. Synaptic weights of the presented stimulus i on trial t are updated according to:

$$G_{t+1,i} = G_{t,i} + \alpha_G G_{t,i}\delta, \tag{2.10}$$

$$N_{t+1,i} = N_{t,i} - \alpha_N N_{t,i}\delta, \tag{2.11}$$

where $\delta = (r_{t,i} - V(s_t))$.

In actor-critic models, actions following positive reward prediction errors are strengthened and more likely to be repeated in the future.

Payoff-Cost model The payoff-cost model is an alternative model grounded in the basal ganglia, which only includes an actor, not a critic. This model accounts for uncertain outcomes and can explain some of the mechanistic effects high or low dopamine levels have on risk preferences. Enhancement of dopamine levels increases overall risk-taking (Gallagher et al., 2007; Orsini et al., 2015; Rigoli et al., 2016; St Onge and Floresco, 2009; Stopper et al., 2014) and risk-preferences are correlated to task-related dopamine activity in the midbrain (Chew et al., 2019).

For example, if a reinforcement takes a positive value r_p half of the times and a negative value $-r_n$ the other half of times, then the Go weights converge to $G = r_p$ and Nogo weights to $N = r_n$, for certain parameters (Möller and Bogacz, 2019). The synaptic weights of the presented stimulus i on trial t are updated according to:

$$G_{t+1,i} = G_{t,i} + \alpha f_\epsilon(\delta) - \lambda G, \tag{2.12}$$

$$N_{t+1,i} = N_{t,i} + \alpha f_\epsilon(-\delta) - \lambda N, \tag{2.13}$$

where

$$f_\epsilon(\delta) = \begin{cases} \delta, & \text{if } \delta > 0, \\ \epsilon\delta, & \text{if } \delta \leq 0. \end{cases}$$

The reward prediction error is $\delta = r - 1/2(G - N)$, where G and N represent the payoff and costs of an action, respectively. The update rules in the above equations consist of two terms. The first term is the change depending on the dopaminergic teaching signal scaled by a learning rate constant α. It increases the weights of Go neurons when $\delta > 0$, and slightly decreases when $\delta < 0$, so that changes in the Go weights mostly depend on positive prediction errors. The constant ϵ controls the magnitude by which the weights are decreased, and takes values within a range $\epsilon \in [0, 1]$, meaning that for $\epsilon = 0$ the negative prediction error does not drive any decrease in Go weights. The details on the values of ϵ required for learning positive and negative consequences are provided in Möller and Bogacz (2019). Nogo weights are updated in a analogous way, but these changes mostly depend on negative prediction errors. The second term in the update rules is a decay term, scaled by a decay rate constant λ. This term is necessary to ensure that the synaptic weights stop growing when they are sufficiently high and allows weights to adapt more rapidly when conditions change. In case an updated weight becomes negative, it is set to zero.

Action selection

Optimal decision-making does not only require learning, but also action selection based on the learned values. When facing a choice between multiple actions, a rational agent that aims to maximise the total future reward chooses the action with the highest (subjective) value and repeats (exploits) this action on subsequent occasions. However, it turns out that humans and animals do not always select the stimulus with the highest probability of reward. Instead, they explore other options occasionally to check whether outcome contingencies have changed. To translate the learned values into choice behaviour, reinforcement learning models capture such choice behaviour with a softmax decision

rule. This rule determines the probability of choosing a stimulus considering the value of the available options. For example, an individual requires to choose between two options in an computerised task and one option is presented on the left (L) and the other one on the right (R). The probability of choosing the left option is computed as follows:

$$P_L = \frac{1}{1 + \exp(-\beta(U_L - U_R))}, \tag{2.14}$$

where U reflects the utility of an action. For Q-learning algorithms, $U = Q$, where all action values are weighted equally. However, U could also represent differently weighted values for economic decisions (chapter 6). The parameter β determines the participant's tendency to exploit (i.e. to choose the stimulus with the highest U value) or to explore (i.e. to randomly choose a stimulus).

Instead of using the general softmax rule, Eq. (2.14), models grounded in the basal ganglia architecture propose a different mechanism through which actions are chosen. The probability of choosing an action depends on the weights in Go and Nogo pathways, which encode positive and negative consequences of actions, respectively. Dopamine controls the relative contribution of these consequences to choice behaviour (chapter 5).

Homeostatic control theories

As discussed in the previous sections, peripheral metabolic signals modulate many neural processes and control elements of behaviour. Homeostasis is often described as a reactive mechanism, but it may be more sophisticated as it requires planning as well (Barrett and Simmons, 2015). For survival and optimal behaviour, individuals need to control and react to changes in physiological state that continuously arise (Friston, 2010; Paulus, 2007). One needs to drive action selection, anticipate consequences of decisions, plan to achieve long-term goals and sometimes override impulses to a maintain physiological balance. The efficient regulation of the body requires the anticipation of phys-

iological needs in order to satisfy them before they arise (Sterling, 2012), a concept known as "allostasis".

Models for the control of homeostasis and the integration of internal and external factors can be divided into two categories based on whether they involve a homeostatic set point or not (Toates, 1981). A homeostatic set-point is a balanced state that represents only a small subset of highly frequented internal states that support survival. Deviations from these states may result in death or reproductive failure. Set-point models detect a physiological state, compare this to the set-point and promote actions when the actual physiological variable falls below the set-point. The ability to be as close as possible to the set-point of internal states that maximise survival, is known as fitness.

Models of motivation

Two theories that describe how incentive processes control the motivation of instrumental action are the drive reduction theory and incentive salience theory. These models do not rely on a homeostatic set-point.

The drive theory (Hull, 1943) argues that primary motivational drives, such as hunger and thirst, simply serve to energise actions. The internal buildup of drive is aversive and its reduction is reinforcing, promoting the repetition of behaviour that causes this reduction. A concept that is better known as the drive-reduction theory. Behaviour is driven by a generalised drive and learned associations. The general drive is used to motivate behaviour in a non-specific way and cannot be used to steer behaviour. To make actions directed towards specific outcomes, Hull included specific drives that are guided by stimulus-response associations learned through reinforcement. Although the drive-reduction theory has provided important insight, it struggles to explain empirical observations, such as eating when sated (Reppucci and Petrovich, 2012) or drinking when hydrated (Gawley et al., 1988). These maladaptive behaviours are caused by drives that occur without the existence of homeostatic errors.

Alternatively, the incentive learning theory (Tolman and Gleitman, 1949) argues that the incentive value, assigned to an outcome produced by an action, modulates whether this action is repeated in the future. The motivational drive does not control the incentive salience directly, but exerts its effects through direct experience with a reinforcer in that state during acquisition. Incentive salience is a specific form of motivation that integrates two types of information: the current physiological state and the previously learned association about reward cues (Berridge, 2012; Berridge and Robinson, 1998). The incentive salience framework (Bindra, 1978; Toates, 1986) postulates a fundamental dissociation in brain mechanisms between the hedonic impact of a reward (reward 'liking') and the motivation for a reward (reward 'wanting'). This model can capture fluctuations in cue-triggered motivation (wanting) as seen in salt appetite and drug-induced intoxication and sensitisation.

Control theory

Early models of homeostasis that involve a homeostatic set-points were adapted from optimal control theory (Carpenter, 2004). These models rely on subsystems with various feedback structures. The sensory system (controller) is combined with the motor system (a plant) to control the current homeostatic state that is sensed by interoceptive input. The current homeostatic state is then compared with the homeostatic set-point and is iteratively minimised by future motor commands. A copy of the motor command is sent to a forward dynamic model, which learns to predict future interoceptive input based on the executed motor commands. Differences between the prediction from the forward dynamic model and the interoceptive feedback create prediction errors, which are used to update the prediction of the forward dynamic model to optimise behaviour.

Homeostatic reinforcement learning

An alternative account for the control of homeostasis, relies on reinforcement learning theory. Reinforcement learning models address how animals adapt to changes in the external environment under the assumption that overall reward is maximised. In homeostatic reinforcement learning, the definition of reinforcements is redefined as outcomes fulfilling physiological needs (Keramati and Gutkin, 2014). These models describe how behaviour changes in response to feedback about the internal state of the animals, assuming that individuals aim to minimise deviations of physiological states from their set-points. Behavioural adaptation to both internal and external states can take place at the same time, because these models integrate learning and regulatory actions. They explain empirical observations, such as motivational sensitivity of different associative mechanisms, anticipatory responses, risk aversion and interaction of competing motivational systems.

Active inference theory

An alternative account for the minimisation of homeostatic prediction errors comes from the active inference theory. Active inference is an approach to understanding behaviour that is based on the idea that the brain uses an internal generative model to predict incoming sensory data. This model establishes a probabilistic mapping between hidden internal or external states and their predicted sensory consequences, exteroceptive or interoceptive inputs, respectively. Prior belief distributions of the interoceptive states represent homeostatic set-points that an agent predicts it will occupy. The internal model combines a prior belief with a likelihood function mapping from hidden internal states to the interoceptive inputs. If discrepancies between the anticipated top-down internal model and interoceptive bottom-up processes occur, prediction errors may arise and the brain tries to minimise these prediction errors. The agent should initiate action to make predictions of the internal model consistent with

perceptual signals. If predictions are inaccurate, self-regulation to achieve long-term goals may be affected as well (Kruschwitz et al., 2019).

2.6 Thesis aims

In this chapter, I reviewed the behavioural, neurobiological and theoretical processes involved in value-based decision-making. Most of the research in human volunteers has focussed on how external factors, such as reward size, delay, and probability, alter decision-making. However, less attention has been directed at understanding how the level of energy reserves modulates learning and action selection. Experimental studies in animals usually rely on food or water deprivation to incentivise animals to work for rewards and learn associations faster. In contrast, human volunteers are usually incentivised by other rewards, such as money. Neurobiological research has shown that food deprivation enhances dopaminergic signalling and increases the incentive salience of food. However, classical reinforcement learning theory only accounts for learning of external factors, which is typically mapped onto dopaminergic function. By contrast, homeostatic reinforcement learning theory accounts for the integration of internal and external factors to modulate decision-making processes, but has not been directly mapped onto the brain's architecture.

To gain insight into how physiological signals interact with the dopaminergic system to influence learning and action selection, my first aim is to extend the reinforcement learning model with a motivational component in a biologically relevant manner. In order to do so, I first need to develop a formal mathematical formulation for how a physiological state affects learning and action selection in the basal ganglia. The prediction error encoded in the dopaminergic activity needs to be redefined to depend on both the objective reinforcement and the motivation. To develop this theory, I am combining influences from economic utility theory, incentive salience theory and build on existing models of learning in the basal ganglia. This model must be able to dynamically

modulate learning and action selection depending on the current physiological state that is consistent with empirical observations.

Empirical evidence suggest that the physiological state is likely to influence both the activation and teaching signals carried by dopamine. Animal studies show that these effects are specific to domain-relevant reinforcers, such a for food when hungry. In my thesis, I explore the possibility that food deprivation could have a generalising effect and also influences the valuation, learning and action selection for domain-irrelevant rewards. This research is divided into four aims that assess the effects of food deprivation on: subjective valuation of food, non-food and monetary outcomes; learning from positive and negative outcomes; risk-taking; and cognitive versus reinforcement-guided action. I combine computerised decision-making tasks with computational models to examine whether food deprivation affects learning and action selection for monetary rewards.

Together, this work provides insight into the mechanisms by which motivational drives, such as hunger, could affect valuation, learning and action selection in a biological manner. It also provides evidence that hunger may have a generalising effects on domain-irrelevant rewards, which could have significant implications for both real-world economic transactions and for aberrant decision-making in eating disorders and obesity.

Experience without theory is blind, but theory without experience is mere intellectual play.

— Immanuel Kant

3

Modelling the effects of motivation on learning and action selection

Contents

3.1 Introduction 39

Effects of motivation on dopaminergic responses 42

3.2 Results 44

Normative theory of state-dependent utility 44

Simulating state-dependent dopaminergic responses 47

Accounting for positive and negative consequences of actions 49

Neural implementation 51

Relationship to experimental data 61

Comparison to an existing incentive salience model 69

3.3 Discussion 72

Experimental predictions 72

Relationship to experimental data and implications 73

Relationship to other computational models 77

3.4 Methods 80

Simulating state-dependent dopaminergic responses 80

Simulating Go and Nogo learning 82

Inductive bias allows for faster learning 82

Influence of physiological state on action selection 83

State-dependent valuation 84

3.1 Introduction

Successful interactions with the environment require the use of previous experience to predict the value of outcomes or consequences of available actions. Human and animal studies have strongly implicated the neurotransmitter dopamine in these learning processes (Bayer and Glimcher, 2005; Frank, 2005; Frank et al., 2004; Pessiglione et al., 2006; Schultz, 1998; Schultz et al., 1997; Steinberg et al., 2013; Wise, 2004), in addition to its roles in shaping behaviour, including motivation (Berridge and Robinson, 1998), vigour (Niv et al., 2007) and behavioural activation (Berridge, 2007; Robbins and Everitt, 2007).

As discussed in chapter 2, dopamine seems to have two distinct effects on the networks it modulates. It is thought to encode reward prediction errors (Schultz et al., 1997), which are used to facilitate learning by triggering synaptic plasticity (Shen et al., 2008), and it controls the activation of the BG network by modulating the excitability of neurons (Hernández-López et al., 2000; Moyer et al., 2007; Thurley et al., 2008). These dopaminergic responses are also known as the teaching and activation signals, respectively.

The overall value of a reinforcement that is available at a given moment depends on the potential positive and negative consequences associated with obtaining it. These consequences can be influenced by internal and external factors, such as the physiology of the individual and the reinforcement's availability, respectively. Dopaminergic responses encode information about the external factors, which scale with the magnitude and the probability of a received reinforcement (Fiorillo et al., 2003; Tobler et al., 2005), but also with the delay and effort related costs associated with a reinforcement (Day et al., 2007; Gan et al., 2009). Internal physiological factors, such as hunger, have also been observed to modulate dopamine levels and dopaminergic responses that encode reward prediction errors (Aitken et al., 2016; Cone et al., 2016; Papageorgiou et al., 2016). Thus, it is likely that the physiological state influences both the

activation and teaching signals carried by dopamine to alter the reinforcement value of an action and drive decision-making based on the usefulness of an action and the outcome at a given time. Strikingly, the physiological state can sometimes even reverse the value of a reinforcement (e.g. salt) from being rewarding in a depleted state to being aversive in a sated state (Berridge and Schulkin, 1989). Moreover, the physiological state during learning may affect subsequent choices. For example, animals may still prefer actions that were previously associated with hunger even when they are sated (Aw et al., 2009). Recent work has proposed how a physiological state can be introduced into reinforcement learning theory to refine the definition of a reinforcement (Keramati and Gutkin, 2014). Despite the importance of the physiological state for describing behaviour and dopaminergic activity, I am not aware of theoretical work that integrates the physiological state into a theory of dopaminergic responses.

The modulation of subjective preference has also been described in the utility theory, which has been extensively used in economics (Stigler, 1950). It is based on the assumption that people can consistently rank their choices depending upon their preferences. Dopaminergic responses also depend on the subjective utility of the obtained reward magnitude, rather than its objective magnitude (Stauffer et al., 2014).

Recent developments indicate that there is a need to extend the general utility function with a motivational component that describes the bias in the evaluation of positive and negative consequences of decisions as a result of changes in the physiological state of an individual. Evidence for this bias comes from devaluation studies in which reinforcements are specifically devalued by pre-feeding or taste aversion. The concept of state-dependent valuation has been studied in various contexts (Dickinson and Balleine, 2002; Levy et al., 2013; Papageorgiou et al., 2016) and in different species, including starlings (Kacelnik and Marsh, 2002; Pompilio and Kacelnik, 2005), locusts (Pompilio et al., 2006) and fish (Aw et al., 2009). These studies suggest that the utility

of outcomes depends on both the (learned) reinforcement value and the physiological state. The incentive salience theory is one of the earliest attempts to capture the relationship between incentive value and internal motivational state (Berridge, 2007).

In this chapter, I aim to provide an explanation for the above effects of physiological state on behaviour and dopaminergic activity with a simple framework that combines incentive learning theory (McClure et al., 2003; Zhang et al., 2009) with models of learning in the basal ganglia. By integrating key concepts from these theories, I define a utility function for actions that can be modulated by internal and external factors. The utility is defined as the change in the desirability of physiological state resulting from taking an action and obtaining a reinforcement. The motivation for a particular resource is defined as the difference between the desired and the current level of this resource, following previous theoretic work (Keramati and Gutkin, 2014).

I do not investigate mechanisms by which teaching and activation signals can be accessed, but I assume that striatal neurons can read out both activation and teaching signals encoded by dopaminergic neurons. Although dopamine is a critical modulator of both learning and activation, it is unclear how it is able to do both given that these processes are conceptually, computationally and behaviourally distinct. For a long time, our understanding was that tonic (sustained) levels of dopamine encode an activation signal and phasic (transient) responses convey a teaching signal (i.e. prediction error) (Niv et al., 2007). However, recent studies have shown that this distinction may not be as clear as initially thought (Feher da Silva and Hare, 2018; Mohebi et al., 2019) and suggest that other mechanisms may exist that allow striatal neurons to correctly decode the two signals from dopaminergic activity (Berke, 2018).

To provide a rationale for my framework the remainder of the introduction reviews the data on the effects of physiological state on dopaminergic teaching signals.

Effects of motivation on dopaminergic responses

I first review a classical reinforcement learning theory and then discuss data that challenges it. As postulated by reinforcement learning theories, expectations of outcomes are updated on the basis of experiences. This updating process may be guided by prediction errors, which are computed by subtracting the cached reinforcement expectation (V_t) from the received reinforcement (r_t). In classical conditioning and after extensive training, the dopaminergic response to the CS reflects the expected future reinforcement, whereas the response to the US represents the difference between the obtained reinforcement and the expectation (Schultz et al., 1997). To account for these responses, the reward prediction error in a temporal difference model (δ_{TD}) is classically defined as (Schultz et al., 1997):

$$\delta_{TD} = r_t + V_{t+1} - V_t. \tag{3.1}$$

The above equation defines the prediction error as the difference between the total reinforcement (including both the received reinforcement r_t and the expected reinforcement in the future V_{t+1}) and the expected reinforcement (V_t).

I now review how the above equation captures the dopaminergic responses at the time of the CS and the US, which change over the course of learning. At the start of learning, the animal has not yet formed any expectations, which means that at the time of the CS, V_t is 0. Given that no reinforcement is provided at the time of CS presentation, r_t is also 0. The prediction error at the time of the CS is equal to the expected value of the reinforcement ($\delta_{TD} = V_{t+1}$), which is zero in naive animals. By contrast, the response to the CS in fully trained animals reflects the expected upcoming reinforcement, because animals have updated their expectations to predict upcoming reinforcements better through extensive experience. At the time of the US, no future reinforcements are expected so V_{t+1} is 0, thus the reward prediction error at the time of US is equal to $\delta_{TD} = r_t - V_t$. Unpredicted rewards (i.e. positive reinforcements)

evoke positive prediction errors, while predicted reinforcements do not. This definition of the prediction error captures observed patterns of dopaminergic responses; naive animals that are unable to predict reinforcements show large positive responses at the time of the US, and fully trained animals that can predict reinforcements perfectly show no response at all.

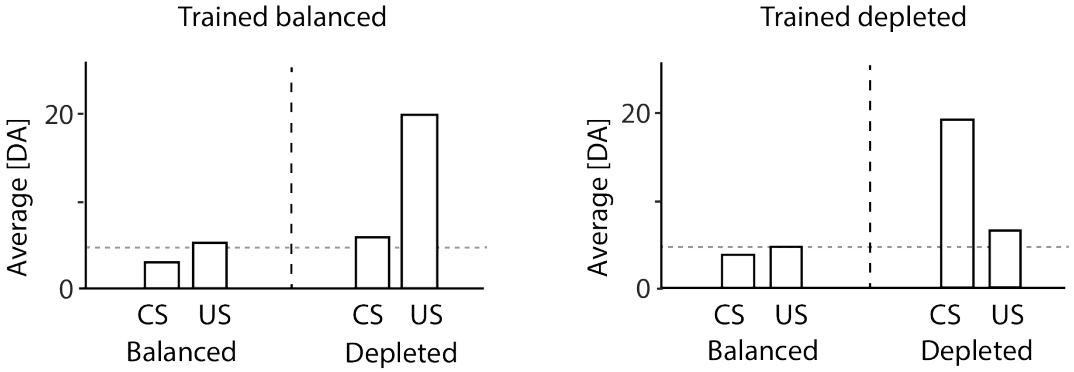


Figure 3.1: **State-dependent modulation of dopaminergic responses**. Experimental data by Cone et al. (2016) shows dynamic changes in dopaminergic responses based on the state of the animal. The two graphs within the figure correspond to dopaminergic responses in animals trained in a balanced and depleted state, respectively, re-plotted from figures 2 and 4 in the chapter by Cone et al. (2016). Within each graph, the left and right halves show the responses of animals tested in balanced and depleted states, respectively. The horizontal dashed lines indicate baseline levels.

Recently, the above definition of reward prediction error has been experimentally challenged by Cone et al. (2016). They showed that the internal state of an animal modulates the teaching signals encoded by dopamine neurons in the midbrain after conditioning (Fig. 3.1). In this study, animals were trained and tested in either a sodium depleted or sodium balanced state. The dopaminergic responses predicted by the classical reinforcement learning theory (Schultz et al., 1997) were only observed in animals that were both trained and tested in the depleted state. In all other conditions, the dopaminergic responses followed different patterns. Dopaminergic responses increased to the US (i.e. salt infusion), rather than the CS, when animals were trained in the balanced state, but tested in the depleted state. This observation is similar to dopaminergic responses observed in untrained animals in the study by Schultz

et al. (1997), suggesting that learning did not occur in the balanced state. No dopaminergic responses to the US or CS were observed in animals that were trained and tested in the balanced state. Interestingly, the same pattern was observed in animals trained in the depleted state and tested in the balanced state, suggesting that the learned values were modulated by the state at testing. In the Results section below, I demonstrate how these patterns of activity can be captured by appropriately modifying a definition of the prediction error.

3.2 Results

In this section, I present the new framework and its possible implementation in the basal ganglia circuit. I also illustrate how it can account for the effects of motivation on neural activity and behaviour.

Normative theory of state-dependent utility

To gain intuition for how actions can be taken based on a state-dependent utility, I first develop a normative theory. The utility and consequences of actions are dependent on the usefulness of the reinforcement (r) with respect to the current state. Taking actions and obtaining reinforcements allows us to maintain our physiological balance by minimising the distance between the current state S and the desired state S^*. I assume that the desirability function of a physiological state has a concave, quadratic shape (Fig. 3.2A), because it is more important to act when you are in a very low physiological state, compared to when you are in a near optimal state. Thus, I define the desirability of a state in the following way (a constant of 1/2 is added for mathematical convenience, as it cancels out in subsequent derivations):

$$Y(S) = -\frac{1}{2}(S - S^*)^2. \tag{3.2}$$

I define the motivation at a given physiological state as the difference between the desired and the current physiological state:

$$m = S^* - S. \tag{3.3}$$

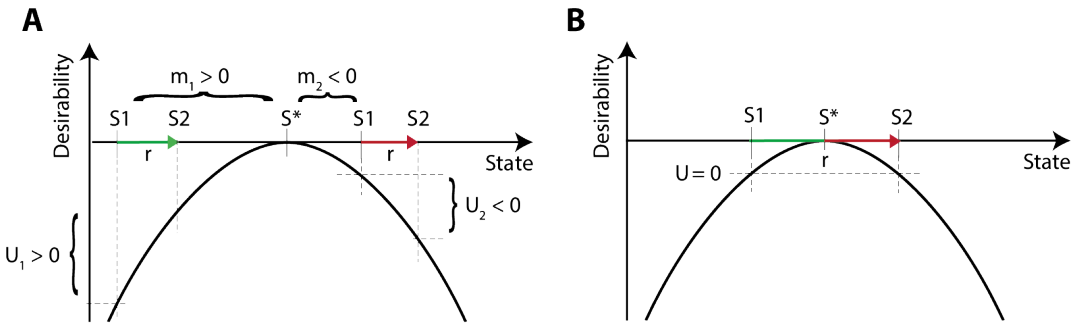


Figure 3.2: **The utility of an action depends on the reinforcement size and physiological state.** A) The same reinforcement can yield a positive or negative utility depending on whether the difference between the current and new physiological state is positive or negative. B) A large reinforcement may have an utility of zero even if the animal was initially in a depleted state. U = utility, m = motivation. S* = the desired state and S1 = state before action, S2 = state after action. Arrow length indicates the size of the reinforcement (r). Changes in state resulting in an increase and decrease in desirability are indicated with green and red arrows, respectively.

For the quadratic desirability function defined in Eq. (3.2), the motivation at any state S is also equal to the slope of the desirability function at this state $Y'(S) = m$. The utility U of an action that changes the state of the animal from $S1$ to $S2$ is defined as the difference between the desirability at state 2, $Y(S2)$ and state 1, $Y(S1)$. A simple approximation for the utility can be obtained through a first order Taylor expansion of Eq. (3.2):

$$\begin{aligned} U &= Y(S2) - Y(S1) && (3.4)\\ &\approx Y'(S1)(S2 - S1) && (3.5)\\ &= mr. && (3.6) \end{aligned}$$

This approximation of the utility clearly shows that the utility of an action, defined as the change in physiological state, depends on both the motivation m (Eq. (3.3)) and the reinforcement r, where $r = S2 - S1$.

According to the above definition, the same reinforcement could yield a positive or negative utility of an action, depending on whether the difference between the current physiological state and the desired state (i.e. the motivation m) is positive or negative (Fig. 3.2A). This parallels an observation that nutrients such as salt may be appetitive or aversive depending on the level of

an animal's reserves (Berridge and Schulkin, 1989). Although not discussed in this chapter, note that this definition of the utility can also be extended to the utility of an external state, such as a particular location in space. The utility of such an external state can be defined as a utility of the best available action in this state.

In order to select actions on the basis of their utility, animals need to maintain an estimate of the utility $\hat{U}$ of an action. There are several ways such an estimate can be learned. Here, I discuss a particular learning algorithm, which results in prediction errors that resemble those observed by Cone et al. (2016). This learning algorithm assumes that animals minimise the absolute error in the prediction of the utility of the chosen action. I can therefore define this prediction error as:

$$\delta = U - \hat{U}. \tag{3.7}$$

The above expression for the prediction error provides a general definition of the prediction error as the difference between the observed and expected utility. I claim that this expression better describes the dopaminergic teaching signal observed in experimental data, and demonstrate this in more detail in the next section.

Assuming that the animal's estimate of expected reinforcement is encoded in a parameter V, the animal's estimate of the utility is $\hat{U} = mV$. Combining Eq. (3.6) with Eq. (3.7), I obtain the following expression for the reward prediction error:

$$\delta = mr - mV. \tag{3.8}$$

A common technique to obtain better predictions is to minimise the absolute error, where the error is defined as the difference between the actual and the predicted utility Eq. (3.7). For this optimisation, I use a gradient ascent and define an objective function F that can be maximised:

$$F = -\frac{1}{2}\delta^2. \tag{3.9}$$

To increase this objective function, the estimate of the expected reinforcement, V, is changed in the direction of the gradient (i.e. the direction that most increases F), which results in an update proportional to the prediction error:

$$\Delta V \sim \frac{\partial F}{\partial V} = m\delta. \tag{3.10}$$

Simulating state-dependent dopaminergic responses

This section serves to illustrate that the pattern of dopaminergic activity seen in the study by Cone et al. (2016) is not consistent with the classical theory and can be better explained with a state-dependent utility as described above. I first simulated the classical model using the reward prediction error described in Eq. (3.1). In the simulation, the CS was presented at time step 1, while the US was presented at time step 2. The model learned a single parameter V that estimates the expected reinforcement on the time step following the CS. Thus, on each trial the prediction error to the CS was equal to $\delta_{TD} = V$, while the prediction error to the US was equal to $\delta_{TD} = r - V$. The value estimate was updated proportionally to the prediction error, i.e. $\Delta V = \alpha(r - V)$, where α is a learning rate parameter. The model received a reinforcement $r = 0.5$ on each trial. Once training was completed, the expected values were fixed to the values they converged to during training, and testing occurred without allowing the model to update its beliefs.

The classical reward prediction error, Eq. (3.1), does not depend on the physiological state, therefore each experimental condition was simulated in exactly the same way. The dopaminergic teaching signals, predicted by the classical theory, were therefore identical in all conditions (Fig. 3.3A). As described above, the response to the CS was equal to the estimate of the reinforcement and the response to the US was equal to the difference between the received and the predicted reinforcement, which was close to zero because reinforcements were fully predicted.

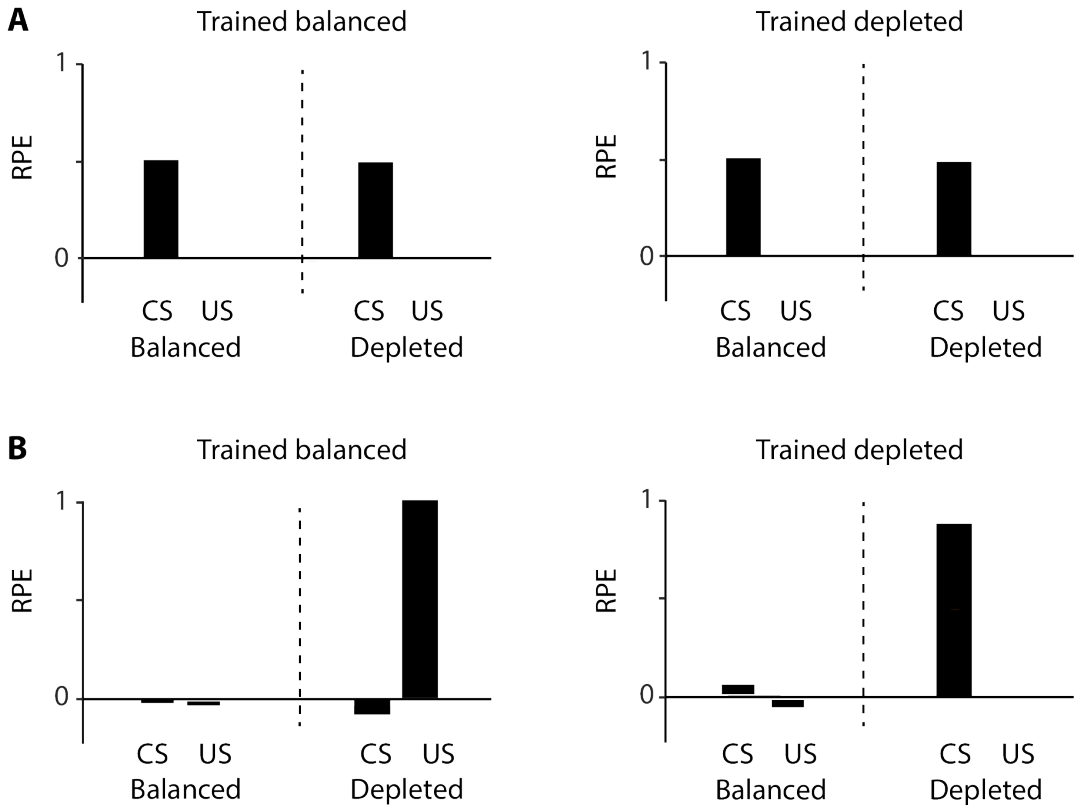


Figure 3.3: **Simulation of data by Cone et al. (2016) with different reward prediction errors.** A) Simulation with classical reward prediction error using Eq. (3.1). B) Simulation with state-dependent prediction error described in Eq. (3.8). Within each graph, the left and right halves show the reward prediction error (RPE) of simulated animals tested in balanced and depleted states, respectively. CS = conditioned stimulus, US = unconditioned stimulus. Each simulation consisted of 50 training trials, 1 test trial and was repeated 5 times, similar to the number of animals in each group in the study by Cone et al. (2016). Error bars are equal to zero as there is no noise added to the simulation and all simulations converged to the same value.

Note that the classical reward prediction error employed in these simulations (Fig. 3.3A) is identical to Eq. (3.8) with $m = 1$. To achieve a state-dependent modulation, I followed the same protocol as in the classical case, but used the state-dependent reward prediction error, Eq. (3.8). The parameter describing motivation was set to $m = 0.2$ for a balanced state and $m = 2$ for a depleted state. Animals trained in the depleted state, exhibited large reward prediction errors to the US in the initial learning phase to facilitate learning, which led to large CS values. By contrast, animals trained in the balanced state exhibited

very small prediction errors, resulting in very small CS values. During testing, the learned CS values were no longer updated. The dopaminergic teaching signal at the time of the US was computed from Eq. (3.8) using the value m of the testing state. The response to the CS was taken as mV, i.e. motivation m at testing multiplied by the CS value learned during training.

Little learning was triggered in simulated animals that were trained and tested in the balanced state, thus the response to the CS was close to zero (Fig. 3.3B). Testing these simulated animals in the depleted state increased the motivation and caused an increase in the scaled utility, which evoked a positive prediction error at the time of the US. By contrast, simulated animals trained in the depleted state learn the estimate of the expected value of the reinforcement. These animals showed an increase in dopaminergic responses to the CS when tested in the same depleted state. However, when these animals were tested in the balanced state the motivation was reduced, the reward prediction error was very small, and little to no dopaminergic responses to the CS or US was evoked as both signals were scaled by a number close to zero.

These simulations show that the prediction error needs to be redefined as the difference between the utility of a reinforcement and the expected utility of that reinforcement, which depends on both the objective reinforcement magnitude and the motivation, to account for the experimental data by Cone et al. (2016). These findings also support the definitions of utility and motivation as described in the previous section.

Accounting for positive and negative consequences of actions

I now reconsider the dependence of utility on motivation and consider how this could be expressed more accurately. Since Eq. (3.6) comes from Taylor expansion, it only provides a close approximation if r is small. This approximation may fail when the reinforcement is greater than the distance to the optimum. Take the example in Fig. 3.2B, if I use a linear approximation with a positive

motivation, the utility is approximated as greater than zero, even though the actual utility is not, because this action exceeds the desired state. Eq. (3.6) also suggests that any action with $r > 0$ has a positive utility if $m > 0$, regardless of possible negative consequences (i.e. reaching a new state further away from the desired state). Moreover, if the distance of the current state to the desired state is equal to the distance of the new state to the desired state, the utility of an action is zero (Fig. 3.2B). It is impossible to capture these effects and account for both positive and negative consequences of this action using Eq. (3.6).

One classical example in which the utility of an action switches sign depending on the proximity to the desired physiological state is salt appetite. Salt consumption is rewarding when animals are depleted of sodium, whereas, salt consumption is extremely aversive when animals are physiologically balanced (Berridge and Schulkin, 1989). To avoid using multiple equations to explain the switch from positive to negative utilities and vice versa (Zhang et al., 2009), I aimed to formulate an equation that can account for negative consequences of actions when the $m \approx 0$ or $m < 0$ and can account for positive consequences when the motivation increases, i.e. $m > 0$.

Therefore, I used a second order Taylor expansion that includes the first and second order derivatives of the desirability function, Eq. (3.2), and gives an exact expression for the utility:

$$\begin{aligned} U &= Y(S2) - Y(S1) && (3.11)\\ &= Y'(S1)(S2 - S1) + Y''(S1)(S2 - S1)^2/2 && (3.12)\\ &= mr - r^2/2. && (3.13) \end{aligned}$$

In the above equation mr could be seen as the positive and $r^2/2$ as the negative consequences of the action. The first term plays a greater role when deprived and promotes taking actions, whereas the second term plays a greater role when balanced and discourages taking actions.

During action selection it is imperative to choose actions that maximise future utility. For competing actions, the utility of all available actions needs to be computed. The action with the highest utility is most beneficial to select, but this action should only be chosen when its utility is positive. If the utility of all actions is negative, no actions should be taken. From a fitness point of view, not making an action may be more advantageous than incurring a high cost.

In the next section, I elaborate on how Eq. (3.13) can be evaluated in the basal ganglia and provide an example of a biologically plausible implementation. For simplicity, I only consider a single physiological dimension (e.g. nutrient reserve), but I recognise that the theory needs to be extended in the future to include multiple dimensions (e.g. water reserve, fatigue) that animals need to optimise. Furthermore, taking an action aimed at restoring one dimension (e.g. nutrient reserve) may also include negative consequences that are independent of the considered dimension (e.g. fatigue). I elaborate on these scenarios in the Discussion.

Neural implementation

In the previous sections, I discussed how the utility of actions or stimuli change in a state-dependent manner. In this section, I focus on the neural implementation of these concepts. More specifically, I address how the utility of previously chosen actions can be computed in the basal ganglia circuit and how it could learn the utility of actions.

Evaluation of utility in the basal ganglia circuit

Given the architecture of the basal ganglia (Fig. 3.4), I hypothesise that this circuitry is well suited for the computation of the utility of actions in decision-making. In particular, I note the similarities between the utility function in Eq. (3.13) consisting of two terms that are differently affected by motivation (i.e. in the first term r is scaled by m, while in the second $-r^2/2$ is not), and the

basal ganglia consisting of two main pathways that are differently modulated by dopamine. Accordingly, the Go and Nogo neurons in the presented model represent the estimates for these two terms and the dopaminergic activation signal encodes motivation and controls the relative influence of Go and Nogo neurons in an appropriate manner. Moreover, I propose that encoding r and $-r^2/2$ separately in the synaptic weights of Go and Nogo neurons, respectively, is beneficial as it allows for asymmetric weighting of the relative contributions of these terms by the dopaminergic activation signal. This mapping may seem counter-intuitive, because the second term of the utility equation is not modulated by motivation, while the Nogo neurons are suppressed by dopaminergic activity. However, in the next few paragraphs, I show that the basal ganglia output encodes a quantity proportional to the utility when dopamine scales the relative effects of the two striatal populations on the output. Furthermore, separating the quantities that are scaled and are not scaled by motivation provides an additional benefit when considering a multi-dimensional space. In the discussion, I highlight some other factors that affect the utility of an action, e.g. related to effort, may be encoded in Nogo neurons but not in Go neurons.

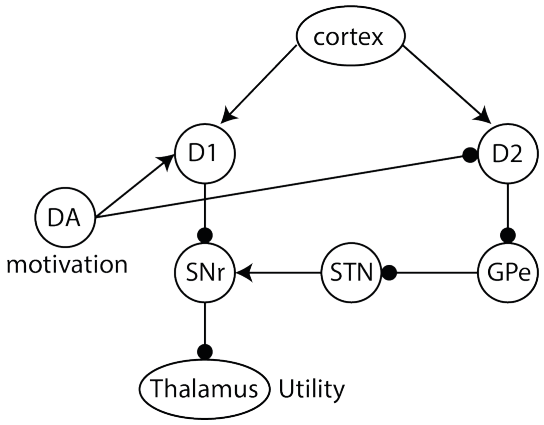


Figure 3.4: **Schematic of utility computation in the basal ganglia network.** Dopaminergic activation signal encodes the motivation. The thalamic activity represents the utility of actions. Arrows and lines with circles denote excitatory and inhibitory connections, respectively. DA = dopamine, D1 = dopamine receptor 1 medium spiny neurons, D2 = dopamine receptor 2 medium spiny neurons, SNr = substantia nigra pars reticulata, STN = subthalamic nucleus, GPe = globus pallidus external segment.

I now consider the computation of the utility at the output of the basal ganglia. I refer to this output as the thalamic activity, denoted by T. T depends on the cortico-striatal weights of Go neurons (G) and Nogo neurons (N), and the dopaminergic activation signal denoted by D.

Although admittedly more complex, the positive and negative contribution of the dopaminergic activation signal, Go and Nogo neurons to the thalamic activity can be captured in a linear approximation (Möller and Bogacz, 2019):

$$T = DG - (1 - D)N. \tag{3.14}$$

In the above equation, the first term DG describes the facilitatory effect of dopamine on Go neurons that results in a positive contribution to the thalamic activity. The second term $-(1 - D)N$ captures the inhibitory effect of dopamine on Nogo neurons, which leads to a negative contribution to the thalamic activity. I assume that $D \in [0, 1]$, and that tonically active dopaminergic neurons give rise to a baseline level of the dopaminergic activation signal of $D = 0.5$ for which both striatal populations equally contribute to the thalamic activity.

In Eq. (3.14), I make three assumptions about the contribution of Go and Nogo neurons to the thalamic activity and how they are modulated by the level of dopamine. For each assumption, I discuss how they relate to experimental data. 1) At baseline dopamine level $D = 0.5$, striatal Go neurons have a positive effect on the thalamic activity of the same magnitude as the negative effect striatal Nogo neurons have on the thalamic activity. This assumption is consistent with an observation that optogenetic activations of Go and Nogo neurons modulate the thalamic activity to the same extent but in opposite direction (see Figure 4 in Lee et al. (2016)). 2) The dopaminergic activation signal increases the gain of Go neurons, and reduces the gain of Nogo neurons. This assumption is based on the observation that dopamine increases the slope of firing-input relationship of neurons expressing D1 receptors (Thurley et al.,

2008), and decreases the slope of neurons expressing D2 receptors (Hernández-López et al., 2000; Moyer et al., 2007). 3) Changes in the level of dopamine affect the gain of Go and Nogo neurons to the same extent. This assumption differs from many models of dopaminergic modulation in which there is an unequal modulation of Go and Nogo neurons based on their low and high receptor affinity, respectively. These models suggests that Go neurons mainly respond to phasic dopamine signals, whereas Nogo neurons mainly respond to tonic dopamine signals (Dreyer et al., 2010). However, my assumption is consistent with recent findings showing that D2 receptors also responds to phasic changes in dopamine, something that is incompatible with the original models (Marcott et al., 2014; Yapo et al., 2017). Moreover, computational models that include both receptor affinity and receptor occupancy show that the effect of dopamine could be similar on Go and Nogo populations (Hunger et al., 2020), which further supports my assumption.

I now show that the thalamic activity defined in Eq. (3.14) is proportional to the utility of an action if G and N are fully learned and therefore provide correct estimates of the positive and negative terms in the utility equation, Eq. (3.13), respectively ($G = r$ and $N = r^2/2$). Then, Eq. (3.14) can be rewritten as:

$$T = Dr - (1 - D)r^2/2 \tag{3.15}$$

which can then be rewritten in the following way:

$$T = (1 - D)\left(\frac{D}{(1 - D)}r - r^2/2\right). \tag{3.16}$$

Comparing Eq. (3.16) to Eq. (3.13), I observe that the thalamic activity is proportional to the utility ($T = (1 - D)U$) when the motivation is encoded by dopaminergic activation signal:

$$m = \frac{D}{1 - D}. \tag{3.17}$$

I can rewrite Eq. (3.17) in the following way to express the level of dopamine for a given motivation:

$$D = \frac{m}{1+m}. \tag{3.18}$$

In summary, when the striatal weights encode the positive and negative consequences, and the dopaminergic activation signal captures motivation, Eq. (3.18), then the thalamic activity is proportional to the utility.

I now consider how this utility can be used to guide action selection. Computational models of action selection typically assume that all basal ganglia nuclei and the thalamus include neurons selective for different actions (Gurney et al., 2001). Therefore, the activity of thalamic neurons selective for specific actions can be determined on the basis of their individual positive and negative consequences and the common dopaminergic activation signal. Given that the proportionality coefficient $(1 - D)$ in Eq. (3.16) is the same for all actions, the utilities of all actions represented by the thalamic activity are scaled by the same constant. This means that the most active thalamic neurons are the ones selective for the action with the highest utility, and hence this action may be chosen through competition. Furthermore, if I assume that actions are only selected when the thalamic activity is above a threshold, then no action will be selected if all actions have insufficient utility. The utility of actions has to be sufficiently high to increase neural firing in the thalamus above the threshold and trigger action initiation. It is important to note that even though the proportionality coefficient $(1 - D)$ in Eq. (3.16) approaches 0 when D approaches 1, this does not mean that the thalamic activity also approaches 0 and no action will be selected. When D approaches 1, the proportionality factor and the utility change concurrently, which means that the thalamic activity either stays the same or increases. More formally, the thalamic activity defined in Eq. (3.14) is non-decreasing with respect to the dopaminergic activation signal, because $\mathrm{d}T/\mathrm{d}D = G + N$ which is non-negative (as synaptic weights G and N cannot be negative).

Furthermore, it may seem counter-intuitive that dopaminergic neurons modulate both Go and Nogo neurons, while the motivation m only scales the first term in the definition of the utility function of Eq. (3.13). The benefit of such double modulation is that it allows for representing an unbounded range of the motivation, $m \in [0, \infty]$, while expressing the dopaminergic activation signal within a bounded range $D \in [0, 1]$. This is useful as biological neurons can only produce finite firing rates. Thus, if an animal is very hungry, setting the dopaminergic activation signal to a value close to a maximum value, denoted here by 1, allows for ignoring any negative consequences of an action. Consequently, the thalamic activity will be positive for any action associated with a positive reinforcement, facilitating the execution of these actions.

Models of learning

In the previous section, I showed that the basal ganglia network can estimate the utility once the striatal weights have acquired the appropriate values. In this section, I address the question of how these values are learned. I already proposed a general framework that describes the learning process under the assumption that the brain minimises a prediction error during this process, and I redefined the prediction error as the difference between utility and expected utility. I described a model in which the thalamic activity encodes a scaled version of the estimated utility ($\hat{U} = T/(1-D)$ (see Eq. (3.16) and (3.17)). This estimate of the utility can be substituted into Eq. (3.7) to give the following state-dependent reward prediction error:

$$\delta = U - \frac{T}{(1-D)}. \tag{3.19}$$

In this reward prediction error, U is the utility of an action (Eq. (3.13)), and $T/(1-D)$ is the expected utility of an action. In line with the general definition of Eq. (3.7), the above equation shows that the reward prediction error depends on a reinforcement and the expected reinforcement in a state-dependent manner, which is here instantiated in a specific model for estimating utility. Note

that at baseline levels of the dopaminergic activation signal ($D = 0.5$), the above expression for prediction error reduces to $\delta = U - (G - N)$ and such a prediction error has been used previously (Mikhael and Bogacz, 2016).

I used two models for learning the synaptic weights of G and N. The first model is a normative model, developed for the purpose of this learning. The second model is the payoff-cost model, which provides a more biologically realistic approximation of the first model. I chose this model as it is able to extract payoffs and costs. Other existing computational models fail to do so and thus I do not expect them to perform optimally (for reference see (Möller and Bogacz, 2019)).

Gradient model The first model I used to describe learning of synaptic weights under changing conditions, directly minimises the error in prediction of the utility of action. It changes the weights of G and N proportionally to the gradient of the objective function: $\Delta G = \alpha \partial F / \partial G$ and $\Delta N = \alpha \partial F / \partial N$, respectively. For the prediction error described in Eq. (3.19), this gives us the following learning rules for G and N:

$$\Delta G = \alpha \delta \frac{D}{1 - D}, \tag{3.20}$$

$$\Delta N = -\alpha \delta. \tag{3.21}$$

Synaptic weights of Go and Nogo neurons are updated using the dopaminergic teaching signal scaled by the learning rate constant α. The update rule for Go weights has an additional term involving the dopaminergic activation signal encoding the motivation as described in Eq. (3.17). In the definition of utility, Eq. (3.13), the motivational level only scales the positive consequences of an action and not the negative, therefore only the update rule for G, but not for N, includes scaling by motivation.

Payoff-cost model The payoff-cost model describes how Go and Nogo neurons learn about payoffs and costs of actions. For example, if a reinforcement

takes a positive value r_p half of the times and a negative value $-r_n$ the other half of times, then the Go weights converge to $G = r_p$ and Nogo weights to $N = r_n$, for certain parameters (Möller and Bogacz, 2019). I expect this learning model to be able to extract positive and negative terms of the utility in Eq. (3.13) if motivation could vary between trials, so the positive term dominates utility on some trials while the negative term on other trials.

The synaptic weights of Go and Nogo neurons are updated using Eq. (2.12) and Eq. (2.13), respectively, using the state-dependent prediction error, Eq. (3.19). The original implementation of this model (Mikhael and Bogacz, 2016; Möller and Bogacz, 2019) uses a reward prediction error that is not state-dependent and is therefore not able to account for decision-making under different physiological states.

Simulations of learning

In this section, I describe under what conditions the learning rules in the gradient and payoff-cost model can yield synaptic weights of Go and Nogo neurons that allow for the estimation of utility. Recall that the network correctly estimates the utility, if $G = r$ and $N = r^2/2$.

I make the following assumptions in my simulations: 1) The simulated animal knows its motivational level m, which influences both dopaminergic signals accordingly (Eqs. (3.18) and (3.19)). 2) The simulated animal computes the utility of an obtained reinforcement as the change in the desirability of the physiological state. As described above, the desirability depends on the objective value of the reinforcement r and the current motivational state m according to Eq. (3.13), which is used to compute the reward prediction error according to Eq. (3.19).

I simulated scenarios in which the simulated animal repeatedly chooses a single action and experiences a particular reinforcement r under different levels of motivation $m \in \{m_{\text{low}}, m_{\text{baseline}}, m_{\text{high}}\}$. Note that $m_{\text{low}} = 0.2$ corresponds to

a dopaminergic activation signal below baseline, $m_{\text{baseline}} = 1$ gives a dopaminergic activation signal of $D = 0.5$, which means that Go and Nogo neurons are equally weighted, and $m_{\text{high}} = 2$ corresponds to a dopaminergic activation signal above baseline levels.

I first simulated a condition in which the motivation changed on each trial, and took a randomly chosen value from a set $\{m_{\text{low}}, m_{\text{baseline}}, m_{\text{high}}\}$ (Fig. 3.5A). The gradient model was able to learn the desired values of Go and Nogo weights as Go weights converged to r, while Nogo weights converged to $r^2/2$, which allowed the network to correctly estimate the utility. Although the subjective reinforcing value changed as a function of physiological state, the model was able to learn the actual reinforcement of an action. Encoding of such objective estimates allows the agent to dynamically modulate behaviour based on metabolic reserves. In contrast, the payoff-cost model converged to lower weights than desired. Although it learned the synaptic weights based on the state-dependent prediction error, the inclusion of weight decay in the model resulted in a lower asymptotic value.

To test robustness of the learning rules and because the motivational state is fixed in the simulated experimental paradigms in this chapter, I also simulated conditions in which the motivation was kept constant (Fig. 3.5B-D). In these cases, both learning rules converged to very similar values of synaptic weights: low levels of motivation emphasised negative consequences and therefore facilitated Nogo learning (Fig. 3.5B), while high levels of motivation emphasised positive consequences and therefore facilitated Go learning (Fig. 3.5D). This shows that when the motivation is equal to the reward size ($m = r$), meaning that this action brings the individual exactly to the desired state S^*, the synaptic weights of Go neurons are higher than Nogo neurons $G > N$. Whenever the action exceeds the desired state, $r > m$, the Nogo weights increase as this action is not favourable $N > G$.

In summary, the simulations indicate that reinforcements must be experienced under varying levels of motivation to learn appropriate values of synaptic

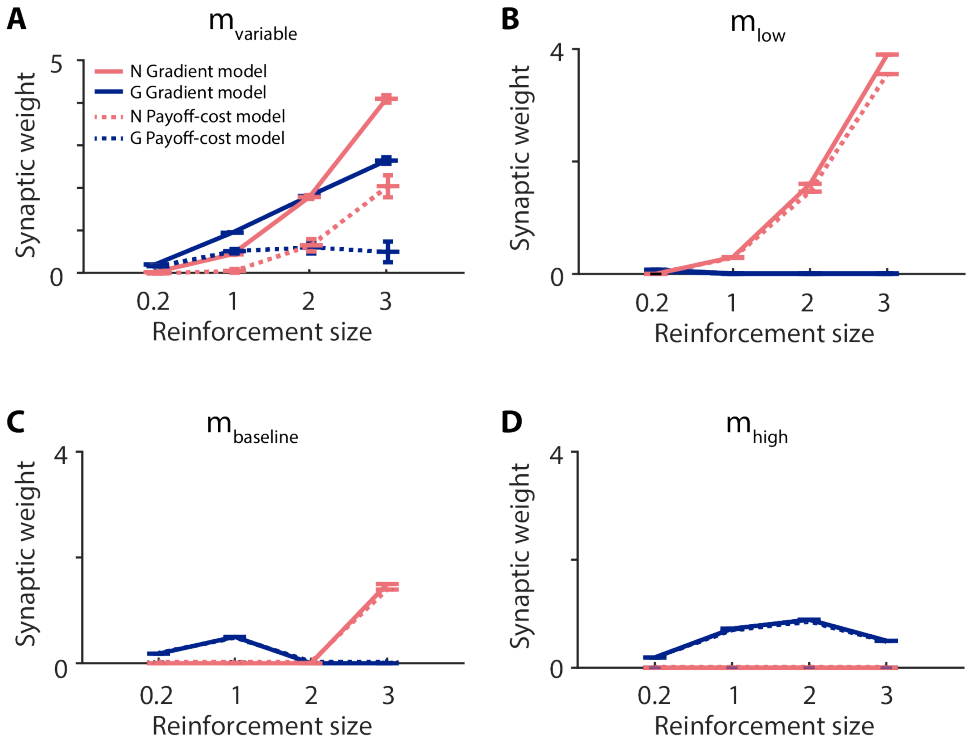


Figure 3.5: **Learning Go and Nogo weights for different reinforcements and different levels of motivation.** Performance of models under variable (A), low (B), baseline (C), high (D) levels motivation. Simulations were performed using the state-dependent prediction error (Eq. (3.19)). Solid lines show simulations of the gradient model using the plasticity rules described in Eq. (3.20) and (3.21). Dashed lines show simulations of the payoff-cost model using the plasticity rules described in Eq. (2.12) and (2.13). Black lines correspond to Nogo neurons and grey lines to Go neurons. Each simulation had 150 trials and was repeated 100 times.

weights. In this case, the gradient model provides a precise estimation, while the payoff-cost model provides an approximation of the utility. In cases when the motivational state is fixed during training, both models learn very similar values of the weights.

The basal ganglia architecture allows for efficient learning

In the previous sections, I presented models and analysed how utilities can be computed and learned in the basal ganglia network. One could ask, why would the brain employ such complicated mechanisms if a simple model could give the same results? In particular, one could consider a standard Q-learning model,

in which the state is augmented by motivation. Such a model would also be able to learn to estimate the utility. However, it does not incorporate any prior knowledge about the form of the utility function and its dependence on motivation. By contrast, the model grounded in the basal ganglia architecture assumes a particular form of the utility function that needs to be learned. In machine learning, such prior assumptions are known as 'inductive bias', and they facilitate learning (Mitchell, 1997).

Thanks to the correct inductive bias, the gradient model learned to estimate the utility faster than standard Q-learning (Fig. 3.6). In my implementation of Q-learning, the range of values the motivation can take was divided into a number of bins, and the model estimated the utility for each bin. On each trial, the model received a reinforcement $r = 1$ and computed the utility using Eq. (3.13), which relied on the current motivation. The Q-value for the current motivation bin was updated by: $\Delta Q_m = \alpha(U - Q_m)$. In Fig. 3.6, I compared a Q-learning approach, in which the motivational state was discretised, with the gradient model in my framework, which did not require discretisation of the motivational state. Both models were able to approximate the utility well, but Q-learning took significantly more trials to do so. Moreover, the more bins were used for the discretisation, the slower learning occurred.

Relationship to experimental data

I already demonstrated how a model employing an approximation of utility can explain data on the effect of physiological state on dopaminergic responses. In this section, I demonstrate how I can use models grounded in the basal ganglia architecture to describe these dopaminergic responses and goal-directed action selection in different experimental paradigms.

I first show that the new, more complex and biologically relevant learning rules can be used to explain the data by Cone et al. (2016). In these simulations, the dopaminergic teaching signal at the time of the CS took on the value of the expected utility ($T/(1 - D)$) and the signal at the time of the

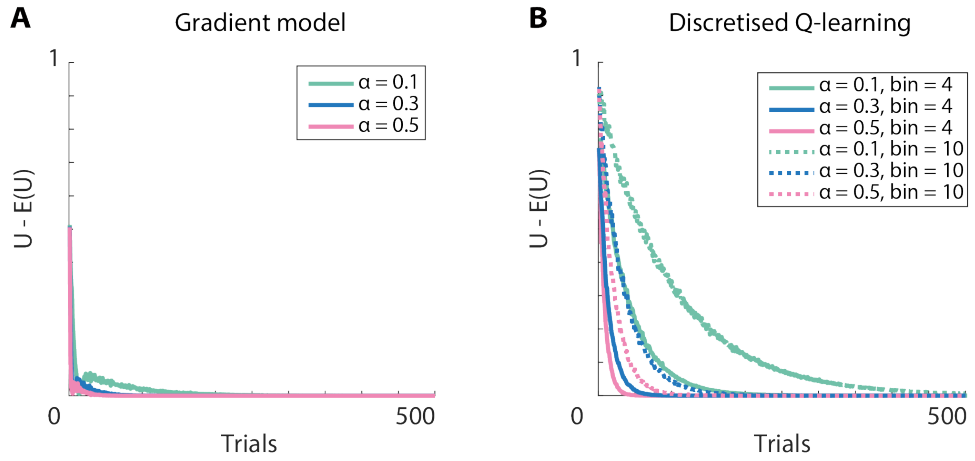


Figure 3.6: **Reward prediction error as a function of learning iteration.** A) The gradient model uses the state-dependent prediction error (Eq. (3.19)) and the plasticity rules described in Eq. (3.20) and (3.21). B) Discretised Q-learning model. Motivational values were randomly chosen on each trial from a uniform distribution between 0 and 2. For Q-learning, motivational values were binned in either 4 or 10 bins. The y-axis corresponds to reward prediction error equal to the difference between the estimated and expected utility.

US represented the reward prediction error described by Eq. (3.19). Simulated values of the dopaminergic teaching signal (Fig. 3.7) showed similar behaviour to the experimental data by Cone et al. (2016). Both the gradient and the payoff-cost model produced a similar dopaminergic teaching signal. This could be expected from simulations in the previous sections, which showed that both models converge to similar weights if the motivation is kept constant during training.

Influence of physiological state on action selection

In the presented framework, natural appetites, such as hunger or thirst, can drive action selection into the direction of the relevant reinforcement. Most food items are generally considered appetitive even when an animal is in the near-optimal state. Nevertheless, overconsumption could have negative consequence as one may experience discomfort after eating too much. Therefore, some of these negative consequences might have to be accounted for as well. As discussed above, a good example of a natural appetite that can be both ap-

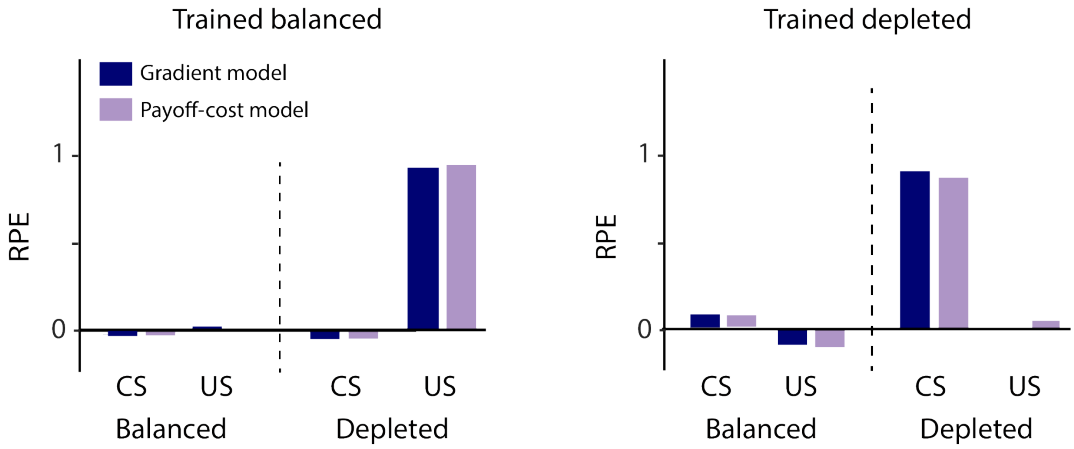


Figure 3.7: **Simulated dopaminergic teaching signal in the paradigm of Cone et al. (2016) according to models grounded in basal ganglia architecture.** For all simulations, the state-dependent prediction error (Eq. (3.19)) was used. The gradient model, depicted in grey, uses the plasticity rules described in Eq. (3.20) and (3.21). Payoff-cost model, depicted in black, uses the plasticity rules described in Eq. (2.12) and (2.13). Left and right panels show the data tested in the balanced state or depleted state, respectively. CS = conditioned stimulus, US = unconditioned stimulus, RPE = reward prediction error. Each simulation consisted of 50 training trials, 1 test trial and was repeated 5 times, similar to the number of animals in each group.

petitive and aversive depending on the physiological state of the animal is salt appetite. Rats reduce their intake of sodium or salt-associated instrumental responding when balanced and vice versa when depleted (Berridge and Schulkin, 1989; Krieckhaus and Wolf, 1968). This even occurs when animals have never experienced the deprived state before and have not had the chance to relearn the incentive value of a salt reinforcement under high motivation (Krieckhaus and Wolf, 1968). This example fits very well with the incentive salience theory, which states that the learned association can be dynamically modulated by the physiological state of the animal. Modulation of incentive salience adaptively guides motivated behaviour to appropriate reinforcements.

I used the study by Berridge and Schulkin (1989) to demonstrate that the simple utility function, Eq. (3.13) can account for the transition of aversive to appetitive reinforcements, and vice versa, in action selection. In this study, animals learned the value of two different conditioned stimuli, one associated

with salt intake (CS+) and one with fructose intake (CS-). The animals were trained in a balanced state of sodium. Once they had learned the appropriate associations, they were tested in a sodium balanced and sodium depleted state. Animals increased their intake of the CS+ in the sodium depleted state compared to the balanced state or the CS- intake (Fig. 3.8A). I assumed that positive and negative consequences were encoded in the Go or Nogo pathway, respectively, and that the synaptic weights of these pathways acquired positive or negative values depending on the situation. Again, the dopaminergic activation signal controlled to what extent these positive and negative consequences affected the basal ganglia output as Go and Nogo neurons were modulated in an opposing manner.

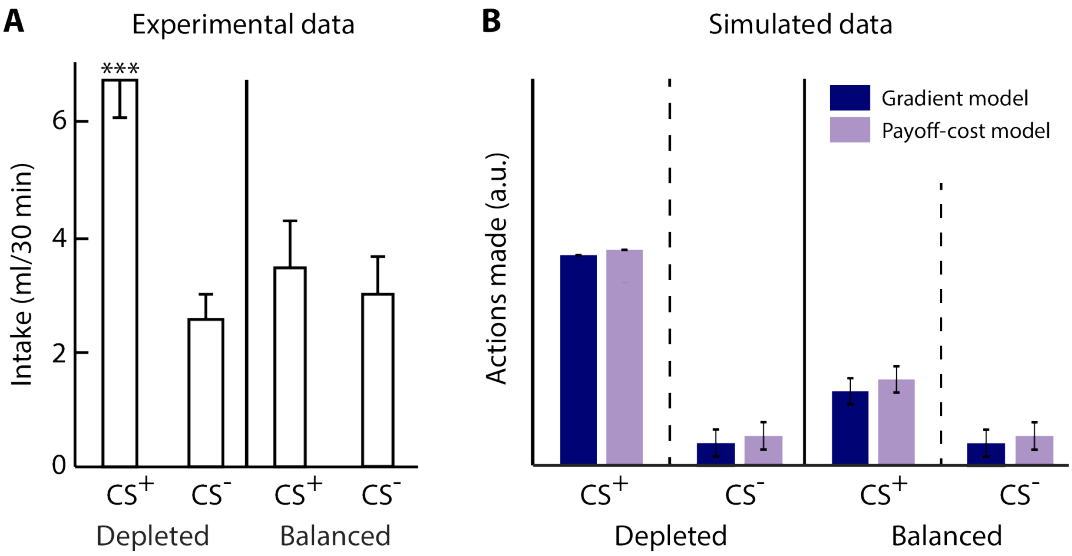


Figure 3.8: **Salt appetite: increased approach behaviour for salt when salt-deprived.** A) Re-plotted data of Berridge and Schulkin (1989) showing that the intake of sodium (CS+), but not fructose (CS-), is increased when salt-deprived. B) Simulated data of number of actions made using the state-dependent prediction error (Eq. (3.19)). The gradient model, depicted in grey, uses the plasticity rules described in Eq. (3.20) and (3.21). The payoff-cost model, depicted in black, uses the plasticity rules described in Eq. (2.12) and (2.13). Within the graph, the left and right halves show the responses of animals tested in depleted and balanced states, respectively. CS+ = relevant conditioned stimulus for sodium, CS- = irrelevant conditioned stimulus for fructose.

Once the appropriate associations between the conditioned stimuli and the outcomes had been acquired, the outcomes were dynamically modulated by the

relevant state only (i.e. the level of sodium depletion). The fact that the responses to the CS- were unaffected by the physiological state of sodium suggests that salt and fructose are modulated by separate appetitive systems and that the physiological state of the animal modulates the intake proportional to the deprivation level of the animal. The phenomenon that different reward types act on different appetitive system has been also observed by other experimental studies (Papageorgiou et al., 2016).

In my simulation, I assumed that the synaptic weights for Go and Nogo neurons were learned in a near-balanced state of sodium since the animals had never experienced a sodium depleted state before. During training, the motivation was low ($m = 0.2$), resulting in a low level of the dopaminergic activation signal (Eq. (3.18)). During the testing phase, the motivation for the CS+ was low ($m = 0.2$) in the near-balanced state and high ($m = 2$) in the sodium depleted state. Given that the experimental data suggest that multiple appetitive systems may be involved. I assumed that sodium physiology did not affect fructose intake and that the animals were not deprived of other nutrients. I used a separate motivational signal for the CS-, which was set to a low, non-negative, value ($m = 0.1$). The thalamic activity was computed using Eq. (3.14), and additional Gaussian noise was added to allow exploration. Actions were selected when the thalamic activity was positive, otherwise no action was selected. The model received a reinforcement of $r = 0.5$ and computed the utility for each action using Eq. (3.13). During training, Go and Nogo weights were updated using the update equations for the different models. For the testing phase, the Go and Nogo values were kept constant based on the learned values and were not allowed to be (re)learned. Again, the thalamic activity was computed and actions were taken when this was positive. Note that the main difference between near-balanced and depleted states, is the level of the dopaminergic activation signal. As shown in Fig. 3.8B, both models show dynamic scaling of the CS+ dependent on the relevant motivational state similar to the experimental data in Fig. 3.8A.

State-dependent valuation

There is a number of experimental studies that have investigated the influence of physiological state at the time of learning on the preference during subsequent encounters (e.g. Aw et al. (2009); Marsh et al. (2004); Pompilio and Kacelnik (2005); Pompilio et al. (2006)). In the study by Aw et al. (2009), animals were trained in both a near-balanced and deprived state. One action was associated with food in the near-balanced state and another action was associated with food in the deprived state. Animals were tested in both states, and in both states, animals preferred the action that was associated with the deprived state during learning. The proportion of trials on which these actions were chosen was above chance level (Fig. 3.9A). These results resemble the data presented in Fig. 3.1, which show an increase in dopaminergic responses to reward-predicting stimuli (CS) when they were experienced in a depleted state. In this section, I show that such preferences can also be produced by the proposed models.

I simulated learning of the synaptic weights of Go and Nogo neurons when the motivation was high (i.e. hungry) and when the motivation was low (i.e. sated). In the experiment by Aw and colleagues, the training phase consisted of forced choice trials in which the reinforcement was only available in one arm of a Y-maze, while the other arm was blocked. For example, the left arm was associated with a food reinforcement during hunger and the right arm was associated with a food reinforcement during the sated condition. In the experiment, 11 animals were used, which were trained for 65 trials on average to reach the required performance. In line with this, my simulations were repeated 11 times, and in each iteration I trained the model for a varying number of trials with a mean of 65. The motivation was set to $m = 2$ for hungry and $m = 0.2$ for sated. The dopaminergic activation signal was fixed to a value that corresponds to the motivation described by Eq. (3.18). For each correct action, the model received a reinforcement of $r = 0.2$ and the utility was calculated using Eq. (3.13). At

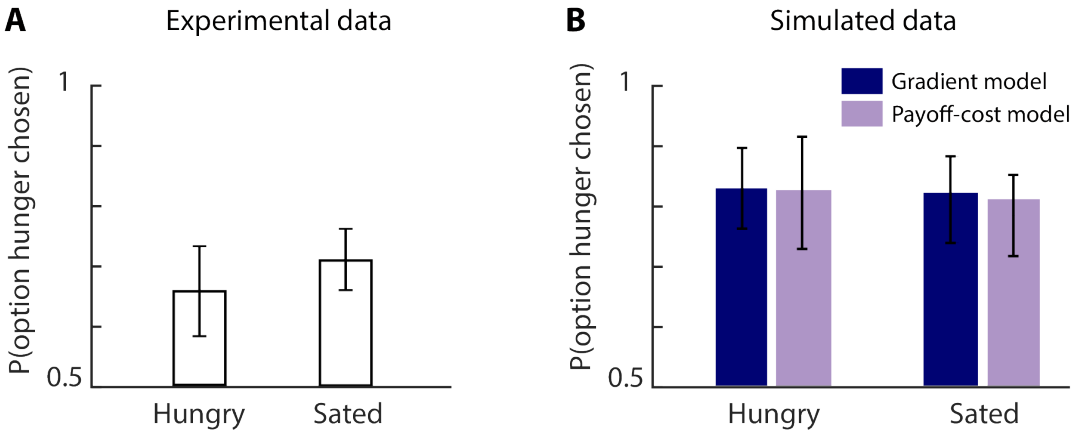


Figure 3.9: **Valuation based on state at training, not testing, predicts preference in choice behaviour.** A) Re-plotted experimental data by Aw et al. (2009) showing that the option associated with hunger during training is always preferred regardless of the state at testing. B) Simulated data using the state-dependent prediction error (Eq. (3.19)). Gradient model, depicted in grey, uses the plasticity rules described in Eq. (3.20) and Eq. (3.21). Payoff-cost model, depicted in black, uses the plasticity rules described in Eq. (2.12) and Eq. (2.13). Hungry and sated refer to the physiological state at testing. The proportion of actions for the arm associated with the hunger condition during training is depicted on the y-axis. For the simulated data this is counted as the number of times the thalamic activity was positive during the test for either the option associated with hunger or sated condition during training.

the start of each simulated forced choice trial, the model computed the thalamic activity of the available action (using Eq. (3.14)) and set the thalamic activity of the unavailable action to zero. Some independent noise was added to the thalamic activity for exploration. The action with the highest positive thalamic activity was chosen. If the thalamic activity of both actions was negative, no action was made and the reinforcement was zero. Each time an action was made the synaptic weights of Go and Nogo neurons were updated using the state-dependent reward prediction error and the update rules described in the section Models of learning. The learning rate for both models was set to $\alpha = 0.1$. Once learning was completed, the synaptic weights were fixed and not updated during the testing phase.

During the testing phase, both arms were available and the animals could freely choose an arm to obtain a reinforcement in. All 11 animals were tested

for 24 trials. The simulated tests were run for 24 trials for both conditions and repeated 11 times using the individually learned Go and Nogo weights for the sated and hungry condition. Again, the model computed the thalamic activity for both options simultaneously (in parallel) plus some independent noise. The action with the highest thalamic activity was chosen. This experiment was simulated for both physiological states during testing. The proportion of actions for the arm associated with hunger were calculated for both states (Fig. 3.9B). Both the experimental and simulated data showed that the animals chose the action associated with the deprived state more often, regardless of the current state.

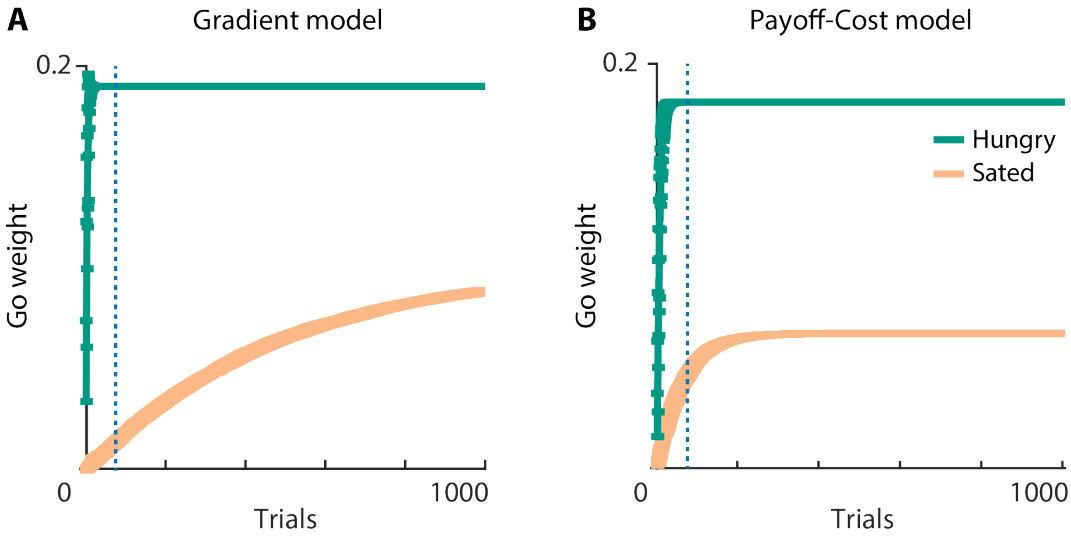


Figure 3.10: **Learning Go weights as function of physiological state.** Go weights were updated using Eq. (3.20) and Eq. (2.12) for the gradient (A) and payoff-cost model (B), respectively. Learning at high motivation is depicted in green and learning at low motivation is depicted in purple. Number of trials used for the simulation was 1000. Go weights were initialised at 0.1.

To gain some intuition for why the models preferred the option that was associated with hungry state during training, let us look back at the simulations presented in Fig. 3.5B-D. These show that when the models were trained with a fixed motivation, Go weights took higher values when the motivation was high during training, and Nogo weights were larger when motivation was low. Analogously, in the simulations of the study of Aw and colleagues, the Go

weights took larger values for the option associated with hunger during training (Fig. 3.10), and this option was therefore preferred during testing. These biases arise in the models when they are unable to experience multiple levels of motivation during training. Only after training in multiple physiological states the utility can be flexibly estimated in different physiological states.

Comparison to an existing incentive salience model

In this section, I formally compare the proposed framework with existing models for incentive salience (Berridge, 2012; Zhang et al., 2009). Zhang et al. developed mathematical models that describe how changes in a physiological state modulate the values of reinforcements. In these models, the learned value of a reinforcement, denoted by R, takes a positive value for appetitive reinforcements and a negative value for aversive reinforcements. A change in the physiological state of the animals from training to testing is captured by a gating parameter κ, which is always non-negative, $\kappa \in [0, \infty)$. An up-shift in physiological state (i.e. when an animal becomes more deprived) is represented by $\kappa > 1$, while a down-shift in physiological state (i.e. when an animal becomes less deprived) is represented by $\kappa < 1$. Up-shifts increase the reward value, whereas down-shifts decrease the reward value. Zhang et al. considered two mechanisms through which a shift in physiological state, κ, modulates the reinforcement, R. According to the first, multiplicative mechanism, the subjective utility of a reinforcement is equal to $R\kappa$. This multiplicative mechanism is similar to the first order approximation of the utility (Eq. (3.6)). However, the model of Zhang et al. assumes that $\kappa > 0$ for all states of the animal, so it struggles to explain a phenomenon such as salt appetite where aversive reinforcements can become appetitive depending on the physiological state of the animal, because multiplying a negative reinforcement with a positive factor does not change the sign of the reward value. To account for this phenomenon, Zhang et al. developed a second, additive mechanism, where the subjective

utility is equal to:

$$U(CS) = R + \log(\kappa) \tag{3.22}$$

For sufficiently large or small κ, this mechanism is able to change the polarity of the utility. As the models of Zhang et al. do not describe learning, it is not possible to compare them explicitly with the simulations of the learning tasks presented in this chapter. Furthermore, these models do not define a reward prediction error, making it impossible to compare their predictions with dopaminergic responses to rewarding stimuli. Nevertheless, I used the additive model (Eq. (3.22)) to compute an utility for all four conditions. I compared these utilities with the dopaminergic responses to the CS in the study by Cone et al., because these responses are thought to reflect the learned values of stimuli. The value of reinforcement was set to $R = -1$ in the balanced state to indicate that salt is aversive in that state, and to $R = 1$ in the depleted state because it is then appetitive. The incentive salience gating parameter was set to $\kappa = 1$ when there was no change between the trained state and the tested state. It was set to $\kappa > 1$ for animals that were trained balanced and tested depleted to indicate an increase in reward value, and set to $\kappa < 1$ for animals that were trained depleted and tested sated to capture a devaluation of the reward. The particular values of κ were chosen to make the utilities qualitatively resemble the dopaminergic responses observed by Cone et al.

Although the model accounted for the dependence of dopaminergic responses on the testing state when animals were trained in the depleted state (Fig. 3.11), the values of conditioned stimuli in the model did not correspond to the observed dopaminergic responses of animals trained in the balanced state (Cone et al., 2016). The dopaminergic responses to the CS were close to baseline for animals trained in the balanced state regardless of the state at testing. The additive model did not produce similar responses, because different values of $\log(\kappa)$ were added in the two testing states. Note that this qualitative difference between the utilities computed by the additive model and

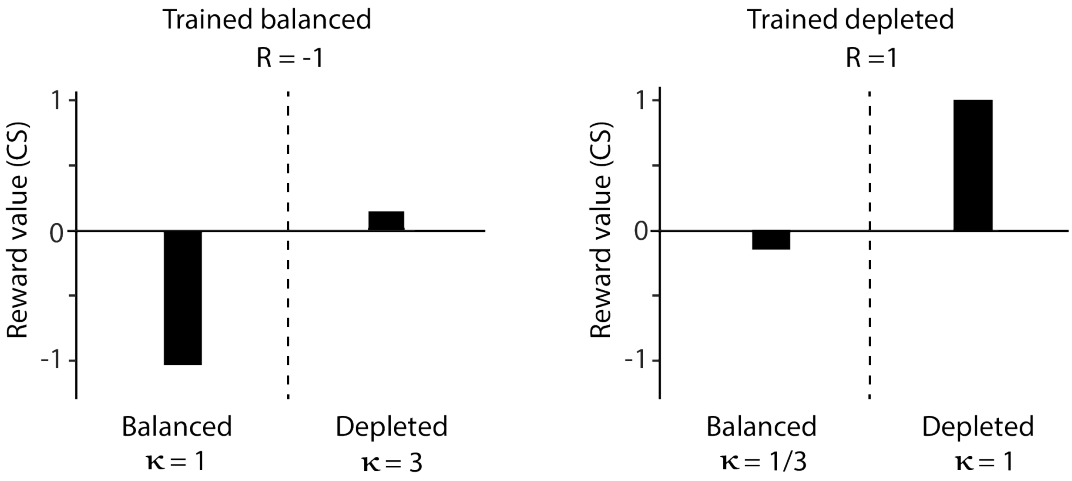


Figure 3.11: **The additive incentive salience model by Zhang et al. (2009) predicts different patterns in dopaminergic teaching signals.** Utility of conditioned stimuli in a study by Cone et al. (2016) was computed according to Eq. (3.22). The reward value was set to $R = -1$ for the animals trained in the balanced state and $R = 1$ for animals trained in the depleted state. The gating parameter $\kappa = 1$ was when the training state is equal to the testing state. An up-shift in state is represented by $\kappa = 3$ and a down-shift in state is represented by $\kappa = 1/3$.

the dopaminergic responses following training in the balanced state remained present, irrespective of the choice of the value of R. Even if I set $R = 0$ to capture that no learning occurred in the balanced state, the utilities of the stimuli remained different (both shifted upwards), because different values of $\log(\kappa)$ were added.

In conclusion, the most obvious difference between my model and the models by Zhang et al. is the form of the utility function. My model relies one mechanism to describe all types of behaviours, whereas Zhang et al. uses two mechanisms to account for different scenarios. Furthermore, my model describes the learning process by which the estimates of a reward value can be learned. Finally, Zhang et al. uses a "gating" parameter, which describes the direction and the size of the shift between the previous state and the new state, whereas the incentive salience in my models is captured by the parameter for motivation m which describes the current state of an animal. The transition itself and the effect it has on action selection and learning is captured by the state-dependent reward prediction error and the thalamic activity.

3.3 Discussion

In this chapter, I have presented a novel framework for action selection under motivational control of internal physiological factors. The novel contribution of this work is that the framework brings together models of direct and indirect pathways of the basal ganglia with the incentive learning theory, and proposes mechanistically how the physiological state can affect learning and valuation in the basal ganglia circuit. I proposed two models that learn about positive and negative aspects of actions utilising a prediction error that is influenced by the current physiological state. In this section, I discuss the experimental predictions, the relationship to experimental data and other computational models and other implications.

Experimental predictions

In this section, I outline the predictions the models make. I hypothesise that the synaptic weights of Go neurons and Nogo neurons are equal to r and $r^2/2$, respectively. Although a relationship between reward and Go/Nogo neuron activity has been demonstrated (Shin et al., 2018), the effects are heterogeneous which makes an exact mapping of G and N difficult. The mapping chosen in this chapter may seem arbitrary, however, from a biological point of view it captures the effects of the sum of Go and Nogo neurons with respect to reward, and the relationship between the thalamic activity and utility with respect to the motivation and the reward that matches the experimental data. Alternative mappings with the same overall effect may be plausible as well and will need to be tested in experimental studies as suggested below. The neural implementation of the framework assumes that Nogo neurons prevent the selection of actions with large reinforcements when the motivation is low. Thus, it predicts that pharmacological manipulations of striatal Nogo neurons through D2 agonist (or antagonist) should decrease (or increase) the animal's

tendency to consume large portions of food or other reinforcers to a larger extent when it is close to satiation, than when it is deprived.

The neural implementation of the framework also assumes that the activity in Go and Nogo pathways is modulated by the dopaminergic activation signal, which depends on motivation. This assumption could be tested by recording activity of Go and Nogo neurons, for example using photometry, while an animal decides whether to consume a reinforcement. The framework predicts that deprivation should scale up responses of Go neurons, and scale down the response of Nogo neurons.

As shown in Fig. 3.5A, the framework predicts that the synaptic weights of Go and Nogo neurons converge to different values depending on the reinforcement magnitude. These predictions can be tested in an experiment equivalent to the simulation in Fig. 3.5A in which mice learn that different cues predict different reinforcement sizes, and experience each cue in variety of motivational states. The weights of the Go and Nogo neurons are likely to be reflected by their neural activity (e.g. measured with photometry) while an animal evaluates a cue at baseline motivation level. I expect both populations to have higher activity for cues predicting higher reinforcements, and additionally, the gradient model predicts that the Go and Nogo neurons should scale their activity with reinforcement magnitude linearly and quadratically, respectively.

Relationship to experimental data and implications

The proposed framework can account for decision-making and learning as a function of physiological state, as shown by the simulations of the data by Cone et al. (2016), Berridge and Schulkin (1989), Aw et al. (2009). More specifically, I proposed that learning occurs based on the difference between the utility and expected utility of an action. This is in line with a study in monkeys that shows that dopaminergic responses reflect a difference in utility of an obtained reward and the expected utility (Stauffer et al., 2014). That study focused on a complementary aspect of subjective valuation of reward, namely that the

utility of different volumes of reward is not equal to the objective volume, but rather to its nonlinear function. In this chapter, I additionally point out that the utility of rewards depends on the physiological state in which they are received.

Furthermore, literature shows that low dopamine levels, as seen in Parkinson's disease patients, drive learning from errors, whereas normal/high dopamine levels emphasise learning from positive consequences (Frank, 2005; Frank et al., 2004; Weismüller et al., 2018). Human imaging studies found that hungry people show an increased BOLD response to high calorie food items, whereas sated people show an increased BOLD response to low calorie food items (Siep et al., 2009). I observe this as well in my simulations: a low dopaminergic activation signal favours small rewards and emphasises the negative consequences of actions encoded in the synaptic weights of Nogo neurons. In contrast, a high dopaminergic activation signal favours large rewards and emphasises the positive consequences of actions encoded in the synaptic weights of Go neurons. However, there are a couple considerations that have to be made with respect to the dopaminergic signal in my simulations. First, I assume that striatal neurons can read out both motivational and teaching signals encoded by dopaminergic neurons (Mohebi et al., 2019). I describe two roles of dopamine neurons, namely activation and teaching signal, but I do not provide a solution to how these different signals are accessed. The function of dopamine neurons is under a current debate and its complexity is not well understood (Berke, 2018). I will leave the details of the mechanisms by which they can be distinguished to future work. I assume that the models, particularly the gradient model, has access to multiple dopaminergic signals simultaneously. Although I recognise that this is a simplified concept of what might be happening in the brain, it still provides new insights into how these different functions affect aspects of decision-making. Further research is necessary to describe the complexity of dopamine neurons in decision-making.

Second, how the prediction error in my model can be computed must be clarified. According to Eq. (3.19), the dopaminergic teaching signal depends on the value of dopaminergic activation signal (D), which scales thalamic input. One possible way in which such a prediction error could be computed, without the dependence of one dopaminergic signal on another, could involve scaling the synaptic input from thalamic axons directly by an input encoding motivational state (m). Indeed, peripheral hormones signalling energy reserves modulate dopamine neurons directly and indirectly via inputs from the hypothalamus. More specifically, the orexigenic hormone ghrelin increases dopamine release (Abizaid et al., 2006), whereas the anorexigenic hormone leptin decreases dopamine release from the midbrain (van der Plasse et al., 2015).

Third, I focused primarily on one dimension, namely nutrient deprivation. However, experimental data suggests that reinforcements are scaled selectively by their physiological needs (Papageorgiou et al., 2016). A nutrient-specific deprivation alters goal-directed behaviour towards the relevant reinforcement, but not the irrelevant one. In contrast, other physiological factors, such as fatigue, may only scale the negative, but not the positive consequences. This hypothesis is supported by data showing that muscular fatigue also alters dopamine levels (Iodice et al., 2017). Together, this suggests that the utility of an action is most likely the sum of all the positive and negative consequences with respect to their physiological needs or other external factors. Therefore, extending the current theory to multiple dimensions is an important direction of future work. In such an extended model, an action that changes the state of multiple physiological dimensions, e.g. hunger, thirst and fatigue, would need to be represented by multiple populations of Go and Nogo neurons. For example, each physiological dimension is encoded by a population of dopamine neurons. The populations that encode hunger and thirst modulate the value of food and drink reinforcements, r_f and r_d, respectively, which are represented by separate populations of Go and Nogo neurons. By contrast, fatigue would modulate effort, which is represented by a population of Nogo neurons. It

would be interesting to investigate if the required number of neurons could be reduced by grouping terms in the utility function that are scaled in a similar way (e.g. $-r_f^2/2 - r_d^2/2$, which are not scaled by any factor in this example). As explained in the Results section, such factors are represented in the model by the Nogo neurons. Future experimental work on how different physiological states modulate individual dopaminergic neurons would be very valuable in constraining such an extended model.

For simplicity, I used the same framework to model experiments involving both classical and operant conditioning, despite their differences in learning. The learning processes in operant conditioning require the execution of an action, while the processes in classical conditioning do not. To capture the difference between these paradigms, reinforcement learning models assume that operant conditioning involves learning action values, while classical conditioning involves learning the values of states, and analysis of neuroimaging data with such models suggests that these processes rely on different subregions of the striatum (O'Doherty et al., 2004). It would be interesting to investigate in more detail how the circuits in the ventral striatum could evaluate the utility of states.

Although the current study does not focus on risk preference, evidence suggest that a link between risk preference and physiological state exists (Levy et al., 2013; Symmonds et al., 2010). Particularly in the payoff-cost model, the dopaminergic activation signal controls the tendency to take risky actions (Mikhael and Bogacz, 2016), thereby predicting that motivational states such as hunger can increase risk-seeking behaviour. The above mentioned studies showed that changes in metabolic state systematically altered economic decision-making for water, food and money and correlated with hormone levels that indicate the current nutrient reserve. Sated individuals became more risk-averse when sated, whereas food deprived individuals became more risk-seeking.

The current study also provides insights into the mechanistic underpinnings of overeating and obesity. Imaging studies using positron emission tomography showed an important involvement of dopamine in normal and pathological food intake in humans. In comparison to healthy controls, pathologically obese individuals show reduced availability of striatal D2 receptors that are inversely associated with the weight of the individual (Stice et al., 2008; Wang et al., 2001, 2009). My theory suggests that the ability to refrain from taking actions and learn from negative consequences of actions, such as overeating, may be diminished when D2 receptors are activated to a lesser extent. The involvement of the dopamine system in reward and reinforcement suggests that a low engagement of Nogo neurons in obese individuals predisposes them to excessive use of food.

Relationship to other computational models

In the Results section, I already briefly compared my model to existing computational models of incentive salience. I addressed some similarities and differences and pointed out how my models could overcome some of the challenges other models struggle with. Here, I relate my framework to several other theories.

The homeostatic reinforcement learning theory, proposed by Keramati and Gutkin (2014), also incorporated physiological states into the reinforcement learning theory. They defined a 'homeostatic space' as a multidimensional metric space in which each dimension represents a physiologically-regulated variable. At each time point the physiological state of an animal can be represented as a point in this space. They also define motivation (to which they refer to as 'drive') as the distance between the current internal state and the desired set point. I built on this theory and included how the brain computes the modulation of learned values by physiology.

For the existence of the desired physiological state, Keramati and Gutkin referred to active inference theory (Friston, 2010). My framework also shares a

conceptual similarity with this theory, in that both action selection and learning can be viewed as the minimisation of surprise. To make this link clearer, I will provide a probabilistic interpretation for action selection and learning processes in my framework. This interpretation is inspired by a model of homeostatic control (Stephan et al., 2016). It assumes that an animal has a prior expectation $P(S)$ of what the physiological state S should be, which is encoded by a normal distribution with mean equal to the desired state S^*. That model assumes that animals have an estimate of their current bodily state S (interoception). It proposes that animals avoid states S that are unlikely according to the prior distribution with a mean S^* (thus they minimise their "interoceptive surprise"), and they wish to find themselves in the states S with high prior probability $P(S)$. Following these assumptions, I can define the desirability of the state as $Y(S) = \ln P(S)$. If I assume for simplicity that $P(S)$ has unit variance and ignore an additive constant, I obtain my definition of a desirability of a state in Eq. (3.2). In my framework, actions are chosen to minimise the surprise of ending up in a new physiological state. The closer this state is to the desired state, the more likely it will be and the smaller the surprise. Furthermore, motivation itself could be viewed as an error in the prediction of the physiological state.

Similar to action selection, animals update the parameters of their internal model (e.g. V, G, N) during learning in order to be less surprised by the outcome of the chosen action. To describe this more formally, let us assume that the animal expects the utility to be normally distributed with mean $\hat{U}$ and variance 1 (for simplicity). Furthermore, assume that during learning the animal minimises the surprise about the observed utility of action U. Therefore, I can define the negative of this surprise as $F = \ln P(U)$. This objective function is equal (ignoring constants) to the objective function defined in Eq. (3.9). Thus in summary, similar to the active inference framework, both action selection and learning could be viewed as processes of minimising prediction errors.

There are many computational models developed for action selection in either very abstract or more biological relevant ways. One of the leading models in describing how dopamine controls the competition between the Go and Nogo pathways during action selection is the Opponent Actor Learning (OpAL) model. This model hypothesises that the Go and Nogo neurons encode the positive and negative consequences of actions, respectively (Collins and Frank, 2014), and high dopamine levels excite Go neurons and low dopamine levels release the inhibition of Nogo neurons. Moreover, existing neurocomputational theories describe how experience modifies striatal plasticity and excitability of the Go and Nogo neurons as a function of reward prediction errors (Collins and Frank, 2014; Gurney et al., 2001; Mikhael and Bogacz, 2016; Shen et al., 2008). In line with action selection models of the basal ganglia, I assumed that the Go neurons encode positive consequences and Nogo neurons encode negative consequences and that the dopaminergic activation signal controls the balance between these neurons. I extended these concepts by combining them with incentive salience theory.

Other models of action selection suggest that the dopaminergic activation signal is associated with an increase in the vigour of actions (Niv et al., 2007). Niv et al. also assumed that the utility of a reinforcement depends on the deprivation level, however, they do not provide a mechanism for how these utilities could be computed and therefore set them arbitrarily. Moreover, they rely on average reward reinforcement learning techniques, which reveal an optimal policy that leads to an average reward rate per time unit. Following this line of thinking, actions with higher utility (i.e. actions taken in a deprived state) cause faster response rates as the opportunity cost of time increases. Although my model does not describe vigour or response times, it could be related to these output statistics thanks to recent work investigating the relationship between activity of a basal ganglia model and the parameters of a diffusion model of response times in a two alternative choice task (Dunovan et al., 2019). This

study showed that a drift parameter of a diffusion model is related to the difference in the activation of Go neurons selective for the two options, while the threshold is related to the total activity of Nogo neurons. In my framework, motivation scales linearly with Go neurons for both options; it enhances the difference in their activity. Based on the data by Dunovan et al. (2019) motivation is expected to increase the drift rate and reduce the threshold, leading to faster responding.

My framework considers for simplicity that all physiological dimensions (e.g. hunger, salt level, body temperature) have unique values with a maximum desirability, and the desirability is lower for both smaller and higher values along the physiological dimension. Although this assumption is realistic for some dimensions, it may not be realistic for the resources animals store outside their bodies. Indeed, according to the classical economic utility theory, humans always wish to have more monetary resources. Although my framework also assumes diminishing utility of larger reinforcements, it differs from the classical theory in that the utility is a non-monotonic function of reinforcement. I also assume that all the resources are directly consumed, which does not allow for the scenario of storing resources. It would be interesting to extend the presented model to include dimensions without finite value maximising desirability.

In conclusion, my modelling framework maps learning of incentive salience onto the basal ganglia circuitry, a circuitry proven to play an important role in action selection. I used key concepts from both lines of theoretical work to develop a framework that is biologically relevant and describes action selection and learning in a state-dependent manner.

3.4 Methods

Simulating state-dependent dopaminergic responses

Here, I give details on the simulations of the state dependent dopaminergic response observed in the study by Cone et al. (2016) using the classical reward

prediction error and the state-dependent reward prediction error. Each simulation consisted of a training phase, in which the expected reinforcements were learned, followed by a testing phase. Each training phase consisted of 50 trials and one testing trial. The simulations were repeated 5 times (similar to the number of animals used in the study by Cone et al.).

In simulations shown in Fig. 3.3, the prediction error at the time of reinforcement was taken as:

$$\delta = mr - mV, \tag{3.23}$$

where

$$m = \begin{cases} 1, & \text{if } TD, \\ 0.2, & \text{if } balanced, \\ 2, & \text{if } depleted. \end{cases}$$

Although temporal difference learning is not dependent on the motivational state and does not include the parameter m, I was able to use Eq. (3.23) with $m = 1$ to simulate it. Payoff to the model was $r = 0.5$. During the training phases, the expected value of the reinforcement was updated using: $\Delta V_t = \alpha\delta$, where $\alpha = 0.1$. During testing trials, the prediction error at the time of unconditioned stimulus was computed from Eq. (3.23), while at the time of conditioned stimulus was taken as mV.

To illustrate that I can also simulate the state-dependent dopaminergic responses with models grounded in the basal ganglia architecture, I also simulated the data by Cone et al. (2016) with the gradient and payoff-cost models (Fig. 3.7). All the parameters were kept the same as described above, but the state-dependent prediction error used at the time of reinforcement was computed from Eq. (3.19). For the gradient model, I used the plasticity rules described in Eq. (3.20) and Eq. (3.21). For the payoff-cost model, I used the plasticity rules described in Eq. (2.12) and Eq. (2.13). The parameters used in the simulations were $\alpha = 0.1$, $\epsilon = 0.8$ and $\lambda = 0.01$.

Simulating learning of Go and Nogo weights

Simulations shown in Fig. 3.5 were run to compare the ability of models to learn the required values of weights with varying levels of motivation and reinforcement magnitudes. Simulations were performed using the state-dependent prediction error, Eq. (3.19). The gradient model used the plasticity rules described in Eq. (3.20) and Eq. (3.21). The payoff-cost model used the plasticity rules described in Eq. (2.12) and Eq. (2.13). I considered 4 different levels of motivation: variable ($m \in [0.2, 1, 2]$), low ($m_{low} = 0.2$), baseline ($m_{baseline} = 1$) and high ($m_{high} = 2$). In the variable motivational condition the model randomly experienced one of the motivation levels, which changed on each trial. For the other three conditions, the motivation was fixed to one value. For each condition there were four possible reinforcement magnitudes $r \in [0.2, 1, 2, 3]$ which were all simulated independently. Each condition was simulated independently and each simulation consisted of 150 trials and was repeated 100 times. All synaptic weights were initialised at 0.1. The parameters used in the simulations were $\alpha = 0.1$, $\epsilon = 0.8$ and $\lambda = 0.01$. These parameters allow the model to converge to positive and negative consequences at baseline motivational state (Möller and Bogacz, 2019).

Inductive bias allows for faster learning

Simulations in Fig. 3.6 show that the gradient model learns the estimate of the utility faster than standard Q-learning as it relies on prior assumption about the form of the utility. The gradient model was simulated using the state-dependent prediction error, Eq. (3.19) and the plasticity rules described in Eq. (3.20) and Eq. (3.21). In my implementation of Q-learning, the range of values the motivation can take was divided into a number of bins. I considered equal sized bins in the range between 0 and 2. To evaluate the influence of discretisation, I compared two bin sizes, namely 4 and 10 bins. The model estimated the utility for each bin Q_m. The Q-value for the current motivation

bin was updated by: $\Delta Q_m = \alpha(U - Q_m)$. On each trial the model received a reinforcement of $r = 1$ and its utility was computed using Eq. (3.6), which relied on the current motivation. The learning rate parameter was also varied across simulations and took values $\alpha \in [0.1, 0.3, 0.5]$.

Influence of physiological state on action selection

In the simulations shown in Fig. 3.8, the relevant and irrelevant conditioned stimuli were learned independently in a training phase. For the relevant cue (CS+, sodium), the motivation was set to $m = 0.2$, and for the irrelevant cue (CS-, fructose) the motivation was set to $m = 0.1$. The motivation was encoded in the level of the dopaminergic activation signal following Eq. (3.18). The training phase consisted of 50 trials and was repeated 5 times. The models used the state-dependent prediction error described in Eq. (3.19), the plasticity rules described in Eq.(3.20) and Eq. (3.21) for the gradient model and the plasticity rules described in Eq. (2.12) and Eq. (2.13) for the payoff-cost model. The thalamic activity was computed using Eq. (3.14), and additional Gaussian noise ($\mu = 0$ and $\sigma = 0.1$) was added to allow exploration. Actions were made when the thalamic activity was positive, otherwise no action was made. The model received a reinforcement of $r = 0.5$ for each action made and the utility was computed using Eq. (3.13). Once the training phase was completed, the Go and Nogo values were fixed to the learned values and not allowed to be (re)learned during the testing phase. During the testing phase, the motivation for the depleted state of salt was set to $m = 2$, for the balanced state to $m = 0.2$ and for fructose to $m = 0.1$ (independent of salt deprivation). Using the different levels of dopaminergic activation signal Eq. (3.18) for the different conditions, I computed the thalamic activity and added a Gaussian noise with $\mu = 0$ and $\sigma = 0.1$. Whenever the activity of an action was positive the action was taken and added to the total number of actions made as depicted in Fig. 3.8.

State-dependent valuation

The simulation shown in Fig. 3.9 was divided into a training and a testing phase. During the training phase, the models learned the values for the option associated with being sated or hungry. These values were learned independently, because only one option was available at the time (forced choice trials). Motivation for sated was set to $m = 0.2$, and for hungry was set to $m = 2$. The models used the state-dependent prediction error described in Eq. (3.19), the plasticity rules described in Eq. (3.20) and Eq. (3.21) for the gradient model and the plasticity rules described in Eq. (2.12) and Eq. (2.13) for the payoff-cost model. Each model learned two Go weights and two Nogo weights. The thalamic activity was computed using Eq. (3.14), and additional Gaussian noise ($\mu = 0$ and $\sigma = 0.1$) was added to allow exploration. The number of trials during the training phase for each simulated animals was chosen randomly from a normal distribution with a mean 65 and standard deviation 5.5 trials, giving similar number of trials to those in the paper by Aw et al. (2009). The animals were tested in either the sated condition or the hungry condition. Motivation for the sated condition, $m = 0.2$, and for the hunger condition, $m = 2$, was used to calculate the corresponding dopamine level using Eq. (3.18). The testing phase consisted of 24 trials during which the individual Go and Nogo weights of option associated with hunger and the sated condition were used to compute a thalamic activity. The training condition that generated the highest thalamic activity was chosen. The simulations (including training and testing) were repeated 11 times (corresponding to the number of animals in the study of Aw et al.). The proportion of actions for the arm associated with the hungry option was calculated (Fig. 3.9). The parameters used in the simulations were $\alpha = 0.1$, $\epsilon = 0.8$ and $\lambda = 0.01$.

Fig. 3.10 was generated using the same method and parameters as described above for the training phase, except the simulations included a 1000 trials and were repeated 100 times.

You know my methods. Apply them.
— Arthur Conan Doyle, The Sign of Four

4
General experimental setup

Contents

4.1 Study design 86
Participants 86
Psychological questionnaires 86
Manipulation of metabolic state 87
Statistics 91
Software 91
Payment 91
4.2 Effects of fasting on hunger, valuation and attention 91
4.3 Computational modelling 92
Maximum likelihood estimation 93
Hierarchical models 93
Parameter recovery 96
Model comparison 96

In this chapter, I describe the methods and study design to test theoretical predictions in humans under the assumptions made in chapter 3. This general experimental setup is used for the experiments described and analysed in the following chapters. I first give an overview of how the metabolic state was manipulated to generate a temporary state of food deprivation. Then, I discuss the control experiments participants performed, and I provide evidence that the manipulation was successful. Finally, I discuss the underlying principles and the statistical methods used for fitting computational models to behavioural data, the validation of the fitting procedures and model comparison.

4.1 Study design

Participants

Thirty-two healthy volunteers (females: 20, mean age: 25.6±6.5) were recruited for this study. All participants were healthy, had no history of psychiatric diagnoses, neurological or metabolic illnesses, and had not used recreational drugs in the past 3 months. All participants had a normal weight (Body Mass Index: 22.9 ± 3.2 kg/m^2), regular eating patterns and no history of eating disorders. Each participant gave written informed consent and the study was conducted in accordance with the guidelines of the University of Oxford ethics committee.

Psychological questionnaires

As part of the screening procedure, participants completed two questionnaires assessing eating beliefs (EBQ18) (Burton et al., 2017) and other eating habits (Bernsmeier et al., 2015), to ensure participants had regular eating habits and did not have a history of eating disorders.

Participants completed a further five questionnaires to account for impulsivity traits and individual differences in reward/punishment sensitivities and

motivation. I used the following questionnaires: Behavioural Inhibition System and Behavioural Activation System (BIS/BAS) questionnaire (Carver and White, 1994) to assess individual differences in sensitivities on subscales of drive, reward and fun seeking. The Sensitivity to Reward and Sensitivity to Punishment Questionnaire (SRSPQ) (Torrubia et al., 2001), which complemented the BIS/BAS questionnaire, and assessed the sensitivity to reward and punishments in a wider context. The Attentional Control Scale (ACS) questionnaire (Derryberry and Reed, 2002) to assess the attentional control in regulating attentional biases related to trait anxiety. The SnaithHamilton Pleasure Scale (SHAPS) questionnaire (Snaith et al., 1995) to assess the hedonic capacity of individuals. And finally, the Lille Apathy Rating Scale (LARS) questionnaire to assess apathy in different domains (Sockeel et al., 2006).

Manipulation of metabolic state

Participants were tested in a fully-counterbalanced, repeated measures design for the effects of food deprivation on various decision-making tasks (Fig. 4.1A). Sessions were approximately 1 week apart (at least 4 days, but no more than 14 days). All sessions took place at the same time of day between 10 am and 1 pm, to minimise time-of-day effects. For one session, participants were asked to refrain from eating and drinking caloric drinks from 8 pm the night prior to testing. For the other session, participants were asked to eat normally the day before and consume a full breakfast within 1 hour of arriving at the lab for testing. The order the tasks were executed in was fixed for all participants (Fig. 4.1C). The effects of fasting on feelings of hunger, valuation and attention are discussed below. The other tasks are discussed in more detail in the following chapters.

Subjective assessment: Rating

I assessed the effect of food deprivation on self-reported feelings of hunger and mental state using a computerised Visual Analogue Scale at the beginning and

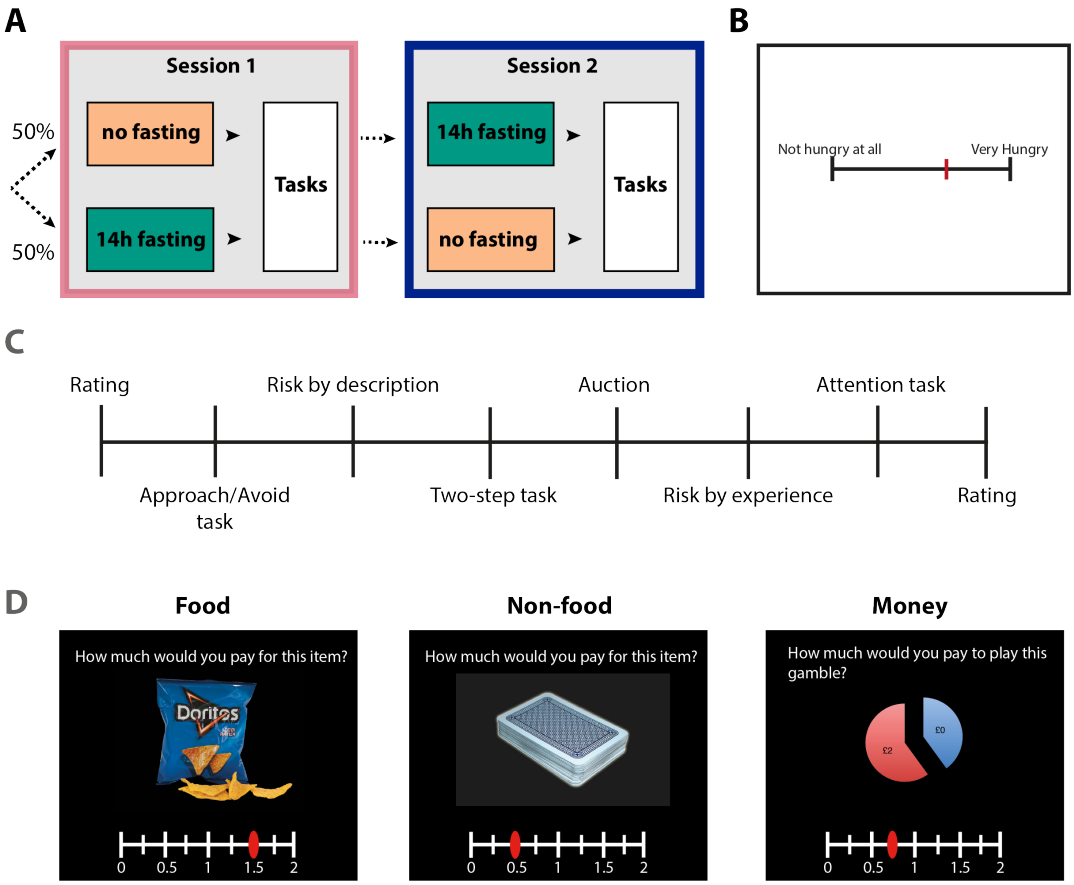


Figure 4.1: **Study design**. A) 32 participants were tested in a fully-counterbalanced repeated measures design. Participants were tested on two separate days approximately 1 week apart. One session took place after 14hours fasting, the other session after consuming a full meal. B) Self-reports of feelings of hunger and other mental states were assessed using a Visual Analogue Scale. C) Timeline and order of performed tasks by participants. D) Assessment of the effects of fasting on valuation were assessed in an auction. Participants were asked to bid an amount between 0 and 2 on items from three categories: food, non-food and monetary gambles. Red dot indicates bid amount (example).

end of each session (Bond and Lader, 1974; Flint et al., 2000). The scale was 100 mm long with positive or negative text ratings anchored at either end (Fig. 4.1B). Participants were asked to place the red cursor wherever they felt was most accurate. This assessment provided a subjective measure of whether the manipulation worked. The following criteria were assessed:

1. Very Hungry / Not hungry at all
2. Alert / Drowsy
3. Calm / Excited
4. Strong / Feeble

5. Clear-Headed / Muzzy
6. Well-coordinated / Clumsy
7. Energetic / Lethargic
8. Contented / Discontented
9. Tranquil / Troubled
10. Quick-witted / Mentally slow
11. Relaxed / Tense
12. Attentive / Dreamy
13. Proficient / Incompetent
14. Happy / Sad
15. Friendly / Antagonistic
16. Interested / Bored
17. Outgoing / Withdrawn

Objective assessment: Valuation

For an objective assessment of food deprivation, I tested participants on their willingness-to-pay for various items. The BeckerDeGrootMarschak method is an incentive-compatible procedure used in experimental economics to measure willingness-to-pay (Becker et al., 1964). This task has been proven to be successful in subjective valuation and identifying corresponding neural substrates (Bushong et al., 2010; Chib et al., 2009; Medic et al., 2014; Plassmann et al., 2007, 2010).

Paradigm Participants were asked to bid on 20 different sweet, savoury and healthy food items (e.g. winegums, doritos, orange), 20 non-food items and 20 gambles for winning either 2 or 4 with a chance of winning ranging from 10-100% (Fig. 4.1C). The items were presented to the participants using high-resolution colour pictures.

In each trial, participants were allowed to bid an amount between 0 and 2 (with increments of 25p) for each item. At the end of the experiment, one of the items they bid on was selected at random. A reference amount was drawn from a uniform distribution between 0 and 2 by pressing a key on the keyboard. If the participant's bid for this item in the bidding phase was higher than the reference amount, the participant won the item or was allowed to play

the monetary gamble. As a result, participants did not have to worry about spreading their 2 over multiple items. They could treat each item as if it were the only decision that counted. The rules of the auction were as follows. Let b denote the bid made by a participant and let n denote the random number used as the reference amount drawn from a known uniform distribution. If $n \leq b$, the participant won the item and paid a price equal to n. In contrast, if $n < b$, the participant did not get the item but also did not have to pay anything. The fact that bidding the willingness-to-pay is the optimal strategy was explained and emphasised extensively during the instructions. I also emphasised that they should not indicate the retail price of an item. The best strategy is to ask themselves how much the item is worth at this point in time, and simply bid that amount (Plassmann et al., 2010).

Attention

Participants were tested on a sustained attention and reaction task (Robertson et al., 1997) to ensure that attention was not affected by the level of food deprivation as this could affect the interpretation of the behavioural results. During this task, participants responded to high-frequency go trials and inhibited their responses to rare nogo trials. Participants were presented with a number between 1–9 in the middle of the screen and were asked to press the space bar for numbers (1,2,3,4,5,6,7,9, "go trials"). When the number 8 was presented on the screen, participants were asked to withhold their response ("nogo trial"). The task consisted of 4 blocks of 30 trials with short breaks in between. The accuracy for both go and nogo trials and the reaction times for go trials were measured.

Statistics

Statistical analyses were conducted using within subject repeated measures analysis of variance (ANOVA) and paired t-tests. Data involving proportions

were transformed using arcsine transformation to ensure data was roughly normally distributed prior to performing statistical comparisons. Correlations were calculated using the Spearmans rank correlation coefficient.

Software

All computerised behavioural paradigms were implemented using Psychophysics Toolbox Version 3 on MATLAB (version 19b, MathWorks, Natick, MA). Statistical analyses were implemented in MATLAB and SPSS (IBM Corp. Released 2019. IBM SPSS statistics for Windows, Version 26.0. Armonk, NY: IBM Corp.).

Payment

All tasks were incentive compatible to ensure that participants chose in accordance with their genuine preferences and maximised their total number of points. All points were converted into pence. Participants were paid 10 per hour and a bonus based on their performance on the tasks. The average time participants spent on the tasks over the two days was between 3.5 - 4.5 hours. Payments ranged from 35 to 45, plus a performance bonus between 0 to 5.

4.2 Effects of fasting on hunger, valuation and attention

I verified whether the manipulation of food deprivation was successful by using a subjective and objective assessment for the level of hunger. Subjective self-reporting showed that participants rated their feelings of hunger significantly higher after 14 hours of fasting than after eating a full meal. (Wilcoxon signed rank test: $[Z = -4.84,\ p < 0.0001]$; Fig. 4.2A). An objective assessment for hunger, through willingness-to-pay, showed that fasting significantly increased the valuation of food $[t(31) = 4.93,\ p < 0.0001]$, but not of non-food items

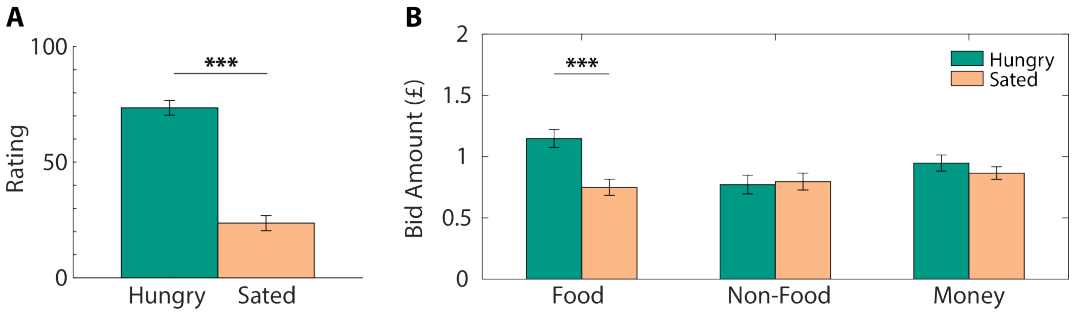


Figure 4.2: **Hunger assessment**. A) Subjective rating of feelings of hunger was significantly increased after 14 hours of fasting. B) Bid amount for food, non-food and monetary gambles. The average amount participants bid was significantly increased for food items after 14 hours of fasting. Error bars are SEM. *** $p < 0.001$.

$[t(31) = -0.47,\ p = 0.642]$ or monetary gambles ($[t(31) = 1.69,\ p = 0.100]$; Fig. 4.2B).

I also tested if food deprivation altered attention by using a sustained attention and reaction task. Participants performed significantly better on go trials compared to nogo trials (main effect of trial type $[F_{1,31} = 77.98,\ p < 0.0001]$). Food deprivation did not affect overall performance (main effect of food deprivation $[F_{1,31} = 1.00,\ p = 0.324]$) or trial-type specific performance (interaction effect of food deprivation and trial type $[F_{1,31} = 1.09,\ p = 0.304]$). Food deprivation significantly increased the median reaction time on the go trials (paired t-test: $[t(31) = 2.77,\ p < 0.009]$). Together, these data showed that food deprivation did not alter attention, but increased deliberation time.

4.3 Computational modelling

To fit the models described in this thesis, I used a maximum likelihood estimation (MLE), which was originally developed by Fisher (Aldrich, 1997). MLE is a common approach used in statistics. I use this technique in chapters 5, 6 and 7 and provide an overview of the method here.

Maximum likelihood estimation

The objective of a model fitting procedure is to identify a set of m parameters $\theta \in \mathbb{R}^m$ that are most likely to have generated the observed data X under the assumptions of a model M. The probability density function defining the chance of observing any given X, given a set of parameters and model, is $P(X \mid \theta, M)$. In MLE, the converse problem is solved; the likelihood of any given θ, given an observed data set X is calculated for which the likelihood function $L(\theta \mid X, M)$ is maximised. This likelihood describes a distribution over possible values of model parameters, rather than a distribution over possible values of the data.

Given that the observed data comprises many trials (samples), the joint probability of the observed data samples, $x_1, \ldots, x_n$, in a data set X, given the probability distribution parameters (θ), can be calculated as $\prod_{j=1}^{n} P(x_j \mid \theta, M)$. Multiplying many small probabilities together can be numerically unstable in practice. This problem is usually restated as the sum of the log conditional probabilities of observing each example given the model parameters $\sum_{j=1}^{n} \log(P(x_j \mid \theta, M))$.

Hierarchical models

MLE finds the best fitting parameters for a dataset of a participant. A hierarchical model fitting approach extends the MLE approach by constraining the estimated parameters by the parameters of the full population (i.e. all participants). The maximum a posteriori estimate for a participant i was:

$$\theta_i = \underset{\theta_i}{\operatorname{argmax}} P(X_i|\theta_i)P(\theta_i|\Theta), \tag{4.1}$$

where $P(X_i|\theta_i)$ is the result of the MLE approach discussed above. The prior distribution on the parameters, $P(\theta_i|\Theta)$, regularises the inference and prevents parameters that are not well-constrained from taking on extreme values as they are normally distributed $\Theta \sim \mathcal{N}(\mu_\Theta, v_\Theta)$, with a population mean of $\mu_\Theta \in \mathbb{R}^m$

and a variance of $v_\Theta \in \mathbb{R}^m$. The parameters of the prior distribution Θ were set to the maximum likelihood, given all the data by all the N participants:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \left(\prod_{i=1}^{N} \int P(X_i|\theta_i) P(\theta_i|\Theta) d\theta_i \right). \tag{4.2}$$

I used an iterative expectation-maximisation (EM) algorithm to compute $\hat{\Theta}$ defined in Eq. (4.2) (Guitart-Masip et al., 2011; Huys et al., 2012; MacKay, 2003). An EM iteration consists of an expectation step to find the best fitting parameters and a maximisation step to update the group level parameters.

On each iteration k, I used Laplacian approximation to estimated the parameter values of each participant by maximising over both the likelihood of participants choices, given the parameters and the likelihood for individual participant parameter values and given the distribution of parameters in the population:

$$\theta_i^k = \underset{\theta_i}{\operatorname{argmax}} P(X_i|\theta_i) P(\theta_i|\Theta^{k-1}) \tag{4.3}$$

In the maximisation step, the population parameters Θ were estimated by setting the mean μ_Θ and the variance v_Θ of the normal distribution to:

$$\mu_\Theta^k = \frac{1}{N} \sum_{i=1}^{N} \theta_i \tag{4.4}$$

$$v_\Theta^k = \frac{1}{N} \sum_{i=1}^{N} \left[(\theta_i^k)^2 + S_i \right] - (\mu_\Theta^k)^2 \tag{4.5}$$

where S_i is the covariance matrix evaluated at the maximum likelihood estimates. S_i was approximated by using the diagonal elements of the inverse Hessian matrix.

The iterative process was stopped when the maximum number of iterations was reached or when the absolute difference in maximum likelihood of the estimated parameters across participants was smaller than a tolerance level ϵ:

$$\left| \sum_{i=1}^{N} \left(L_i^{k+1} - L_i^k \right) \right| < \epsilon, \tag{4.6}$$

where $L_i^k = L(\theta_i^k \mid X_i, M)$ and $\epsilon = 10^{-3}$. The maximum number of iterations was set to 20. For iteration $k = 1$, the mean of the population parameter distribution was initialised at random and the variance was set to $v = 100$. The maximum a posteriori parameter distribution obtained from the model fit across testing sessions was used as the prior for when the parameters of the two sessions were estimated separately.

I used the MATLAB optimisation function *fminunc* to obtain unconstrained parameter estimates. In most of my models, parameters were bounded to a certain range to ensure that parameters were biologically plausible and interpretable. Before inference, these parameters were suitably transformed to enforce the constraints using logistic, exponential or hyperbolic tangent functions. I transformed a [0,1]-bounded parameter, denoted α, into a Gaussian scale using the logistic function:

$$\alpha = 1/(1 + \exp(-a)), \tag{4.7}$$

a [-1,1]-bounded parameter, denoted γ, into a Gaussian scale using the hyperbolic tangent function:

$$\gamma = \frac{\exp(g) - \exp(-g)}{\exp(g) + \exp(-g)}, \tag{4.8}$$

and a $[0, \infty)$-bounded parameter, denoted β, was logarithmically scaled using the exponential function:

$$\beta = \exp(b). \tag{4.9}$$

The model parameters are denoted by Greek letters and their respective Gaussian transformations by Latin letters. Normally distributed parameters allow for the use of parametric tests to identify differences between sessions.

Gradient-based optimisation algorithms are not guaranteed to converge to a global maximum and could instead converge to local maxima. To avoid this,

I initialised the algorithm with a range of starting parameters and chose the iteration with the highest likelihood value to make further inference. All model fitting procedures were verified on surrogate data generated from a known decision process.

Parameter recovery

To validate the parameter estimates generated by the fitting procedure, I conducted a parameter recovery analysis. For each parameter, I randomly generated values using the posterior distribution of the fit to get realistic parameters that could describe choice behaviour. The generated parameters were uncorrelated, allowing me to test whether the fitting procedure introduced any confounding factors. I used the generated sets of parameters to simulate choice behaviour in a decision-making task. Next, I used a hierarchical model fitting procedure to estimate parameters for the simulated data, which I refer to as recovered parameters. I assessed the quality of the parameter recovery by comparing the true parameters used to simulate data to the recovered parameters. I calculated the Pearson correlation between parameters. A strong correlation between the true and recovered parameters indicates a good recovery of the parameters and reliable model fitting results.

Model comparison

To test alternative hypotheses about behaviour and the underlying computational mechanisms, one can design and compare various models that define different mathematical processes. To identify the best fitting model one can compare the maximum likelihood estimates or residual values of all models, where the model with the highest likelihood is most likely to describe the data the best. However, such a comparison does not account for model complexity, i.e. the number of parameters included in the model. An over-parametrised model may fit the data well, but may be a poor out-of-sample predictor compared to a simpler model. I use the Bayesian Information Criterion (BIC) value

to compare models with different complexities, because it modifies the likelihood by penalising models with a large number of parameters (Schwarz, 1978). These modified likelihood values can be compared to identify the best-fitting model, where the lowest BIC value indicates the winning model. I based the model comparison on the BIC values of the population data across conditions for each model.

Good decisions come from experience, and experience comes from bad decisions.

— Mark Twain

5

The effects of metabolic state on approach-avoidance behaviour

Contents

5.1 Introduction 99

5.2 Methods 102

Modified approach-avoidance learning paradigm 102

Behavioural Analyses 105

5.3 Results 105

Theoretical predictions for approach-avoidance learning . . 105

Food deprivation attenuated avoidance behaviour 111

Food deprivation did not attenuate learned values 114

5.4 Discussion 121

Comparison of computational models 122

Effect on learning 124

Effect on action selection 125

Neural substrates 126

Limitations 128

Conclusions 128

5.1 Introduction

For survival, individuals should aim to maximise reward via approach behaviour and minimise punishment via avoidance behaviour. In chapter 3, I described how learning of positive and negative consequences of actions can be mapped onto the basal ganglia circuitry. I argued that learning and action selection are under the direct control of the current physiological state of an individual encoded in the level of dopamine. Using my theory, I predicted that hunger enhances the bias for approach behaviour, whereas satiety increases the bias for avoidance behaviour. In this chapter, I aim to answer two questions: 1) Does food deprivation modulate approach and avoidance behaviour by altering how values are learned or by how values are used to drive choice selection? 2) Are the biasing effects on behaviour driven by changes in sensitivity to positive and negative reward prediction errors or by reward valence?

A frequently used paradigm to test approach and avoidance learning is the probabilistic selection task developed by Frank et al. (2004). This task has given valuable insight into how low dopamine levels, as seen in PD patients, promote avoidance behaviour via increased D2-receptor activity in the Nogo pathway of the basal ganglia. High dopamine levels, caused by dopamine gene mutations or dopamine medication, reverse this bias by activating the D1 receptor in the Go pathway (Cox et al., 2015; Frank et al., 2007, 2004; Jocham et al., 2011; Klein et al., 2007). A central aspect of this task is that participants first learn reward probabilities (i.e. the frequency of good and bad outcomes) associated with different stimuli through reinforcement learning, followed by a testing phase. In the testing phase, individuals guide decisions for novel combinations based on the associated stimulus values previously learned (Frank et al., 2004). The interpretation of test performance is as follows: individuals who are better at approach behaviour select symbols previously associated with frequent good outcomes. Individuals who are better at avoidance behaviour more often reject symbols that were previously associated with frequent bad outcomes.

Mechanistic models of the basal ganglia showed that a bias for approach behaviour may be driven by an increased learning rate for positive reward prediction errors during the reinforcement learning phase or an increased weighting of Go neurons, which encode positive outcomes. In contrast, a bias for avoidance behaviour may be driven by an increased learning rate for negative reward prediction errors or an increased weighting of Nogo neurons, which encode negative outcomes (Collins and Frank, 2014; Mikhael and Bogacz, 2016). Empirical studies have shown that dopamine affects choice behaviour during the test, but not reinforcement learning (Shiner et al., 2012). Dopamine did not alter the learned values at the end of training, even though it modulated the learning rates for positive and negative feedback in opposite direction (Aberg et al., 2016; Frank et al., 2007, 2004; Jocham et al., 2011; Klein et al., 2007). These data suggest that all individuals were able to learn to a similar extent with sufficient training time, regardless of differences in the integration of positive and negative feedback.

Positive and negative reward prediction errors are critical for learning and occur regardless of whether the reward value has a positive or negative valence. This is important as the majority of the studies employing a probabilistic selection task provided abstract feedback, which does not have an inherent reward value. The feedback consisted of either the words correct/incorrect or a happy/sad face to signify a good or bad outcome, respectively. Although abstract feedback may be interpreted as positive or negative, it is unclear whether the brain assigns a reward value to abstract feedback and whether this is equal in all individuals. In particular, incorrect feedback could be perceived as a reward with negative valence, or a reward of zero value. Both interpretations cause prediction errors during learning, but the integration of non-negative outcomes may recruit different brain areas compared to outcomes with negative valence. Many empirical studies assigned non-negative outcomes to good and bad feedback (Frank et al., 2007; Grogan et al., 2017; Jocham et al., 2011),

whereas computational studies assigned a mixture of both positive and negative valence to feedback (Collins and Frank, 2014; Mikhael and Bogacz, 2016). Consequently, learning about outcomes with positive or negative valence, rather than positive or negative prediction errors, generate different predictions about approach and avoidance behaviour than previously reported.

To allow for the comparison with previous studies, while including monetary feedback instead of abstract feedback, I first simulated behaviour using computational models, grounded in the basal ganglia architecture, to guide the design of this study. The introduction of a gain domain (positive valence) and loss domain (negative valence) provided the strongest predictions for differentiating between learning from positive and negative reward prediction errors and learning from positive and negative valence. It also allowed me to differentiate between computational models. If there is an effect on the sensitivity to positive or negative prediction errors, I expect to find a difference between approach and avoidance behaviour regardless of the valence of the outcome. If there is an effect on outcome valence, I expect that the overall performance for choosing and avoiding positive valence outcomes differs from the overall performance for choosing and avoiding negative valence outcomes.

Following chapter 3, I assumed that the level of energy reserves is encoded in the level of dopamine. I further assumed that only domain-specific rewards are valued more, which would be reflected by increased learning rates for these rewards when selectively deprived. In this task, participants received a monetary reward, not a food reward. Therefore, I did not expect the perceived utility of these rewards to be affected. The validity of this assumption was confirmed by the valuation data (Fig. 4.2). This group of participants showed increased valuation for domain-specific food rewards, but not for domain-irrelevant non-food or monetary rewards. This suggests that the learned values in this study will likely depend on the reward probabilities and reward magnitude of the stimuli, but not on the level of hunger.

5.2 Methods

Modified approach-avoidance learning paradigm

The original probabilistic selection task was modified to distinguish learning from positive and negative feedback, and learning in appetitive and aversive domains. I used geometrical shapes instead of the Japanese characters that were used in the original design to increase stimulus discriminability (Schutte et al., 2017). I changed the abstract feedback, "correct" or "incorrect", into quantifiable outcomes (i.e. win or lose 0.50 or 0) to have a clear mapping of reward and allow for the investigation of valence effects. I provided feedback for the chosen and unchosen stimulus (factual and counterfactual feedback, respectively) to facilitate learning. I also introduced a gain and loss domain that were identical in terms of outcome probabilities, but differed in valence. Thus, differences in performance between the gain and loss domain could be attributed to a bias that reflects the difference in sensitivity to positive or negative valence.

The modified task structure was as follows. I used four different stimulus pairs (AB, CD, EF, GH) that were associated with different probabilities to win 0.50 (i.e. gain domain) or lose 0.50 (i.e. loss domain) (Fig. 5.1A). These pairs were categorised based on difficulty, where 80/20 pairs were easier to differentiate and 60/40 pairs were harder to differentiate.

Training phase The training phase consisted of 320 trials, divided into four blocks of 80 trials with a short break after each block. Each stimulus pair was presented an equal number of times in each block. The trial order was pseudo-randomised across participants, but kept constant for both conditions to avoid potential order effects. Stimulus sets were counterbalanced across participants and different stimulus sets were used for the two sessions. The outcomes were generated prior to the task according to the associated probabilities for winning and losing a monetary amount of 0 or 0.50 to ensure that each participant

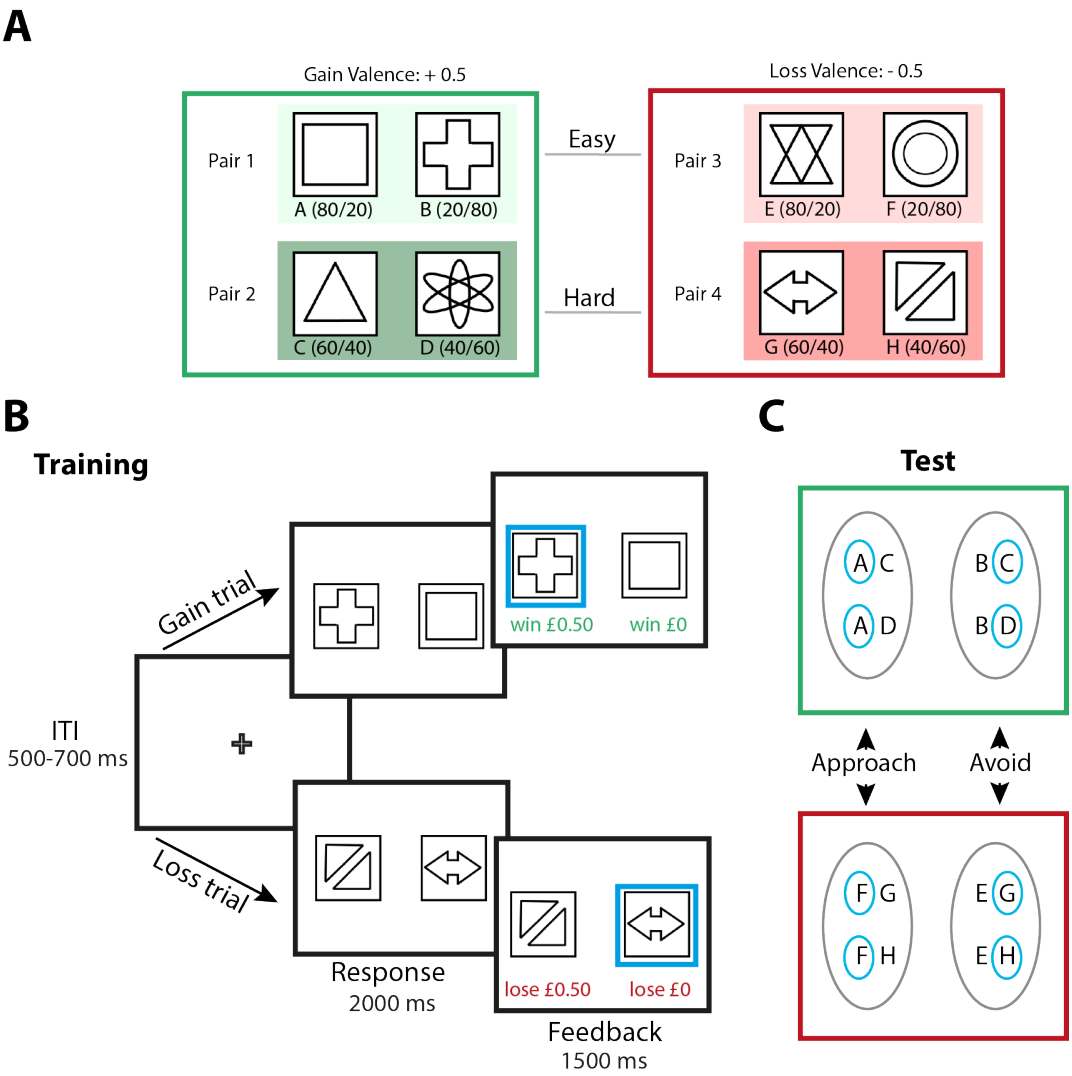


Figure 5.1: **Task design.** A) Four different reward probabilities were used, (80%, 60%, 40%, 20%), for both the gain and loss domain. The stimuli in two pairs had a large difference in reward probabilities (pair 1 and 3, reflected by the light shading) and two pairs had a small difference in reward probabilities (pair 2 and 4, reflected by darker shading). The stimuli pairs follow a two-by-two design: gain/easy (pair 1), gain/hard (pair 2), loss/easy (pair 3), loss/hard (pair 4). The valence of gain trials was +0.50, the valence of loss trials was on average −0.50. The symbols associated with each reward probability were randomised across participants and sessions. B) Example training trial in the gain domain and loss domain. After fixation, two symbols were presented and participants selected one symbol within 2 seconds. Immediately after the choice was made, factual and counterfactual feedback was presented based on the reward probability associated with the presented symbols. C) During the test, novel combinations of stimuli were presented. Participants had to choose the stimulus with highest expected value (A or F) and avoid the stimulus with the lowest expected value (B or E). The correct stimulus is circled in blue.

experienced the intended reward probability. The outcome order was shuffled across participants, but kept constant for the food deprived and sated condition for each participant.

Each trial had the same structure (Fig. 5.1B). After a short inter-trial-interval (ITI) of 500-700 milliseconds, one of the stimulus pairs was presented on the screen. Participants needed to respond within 2 seconds by pressing on the left or right arrow key of the keyboard to choose the left or right option, respectively. The choice was highlighted by a blue square. After each choice, visual feedback was immediately provided for 1.5 seconds, consisting of "win 0.50 or "win 0 in green for the gain domain and "lose 0.50 or "lose 0 in red for the loss domain. If no response was made within two seconds, the words Failed to respond in time were printed in the middle of the screen and no feedback was provided. Three participants were excluded because they failed to select the most rewarding stimulus (A) over the most punishing stimulus (B) and (F) over (E) in the 80/20 pairs on more than 60% of the training trails or more than 50% of the test trials. Data from participants who failed these criteria are not interpretable (Cavanagh et al., 2011; Frank, 2005; Frank et al., 2007). Eight participants were excluded because they did not complete the full task.

Testing phase After the training phase, participants were tested on novel combinations of stimulus pairs involving A (AC and AD) or B (BD and BC) in the gain domain and F (FG and FH) or E (EG and EH) in the loss domain (Fig. 5.1C). In these novel pairs, participants were asked to choose the stimulus with the highest expected value and avoid the stimulus with the lowest expected values, meaning that they had to choose either A or F and avoid B or E (Fig. 5.1C). Participants were also tested on the original pairs encountered during training to verify whether they were able to rely on what they had learned. No feedback was provided during the test to avoid relearning of stimulus outcomes. Participants were instructed to rely on their "gut-instinct" if they were unsure. The response time for each trial was limited to 2 seconds.

Trials without a response or a response time shorter than 200 ms were removed from further analyses. The test consisted of 40 trials, of which eight trials were allocated to the original pairs (two trials per pair), and the remaining 32 trials were used for novel combinations of stimuli (four trials per test pair).

Behavioural Analyses

My primary measure was the proportion of correct choices made on test trials during each session by every participant. In the training phase, participants learned that A is better than B, which could be achieved by either learning that choosing A leads to a positive outcome, or that choosing B leads to a negative outcome (or both). Approach behaviour was quantified as the accuracy of choosing the most optimal stimulus (A and F), which suggests that people learned more from positive feedback. Avoidance behaviour was quantified by the accuracy of avoiding the most punishing stimulus (B and E), which suggests that individuals learned more from negative feedback. This accuracy measure provided an indication of any consistent changes between deprivation states across participants. The secondary measure was the accuracy during the training phase to verify whether differences in learning contributed to effects measured during the test. Performance during training was explained using reinforcement learning models.

5.3 Results

Theoretical predictions for approach-avoidance learning

I first examined what predictions the models generate for the modified version of the task, under the assumption that food deprivation increases the dopaminergic motivation signal (chapter 3). I simulated the choice behaviour of 100 virtual participants on this task using the OpAL (Collins and Frank, 2014) and the payoff-cost model (Mikhael and Bogacz, 2016; Möller and Bogacz, 2019). These reinforcement learning models have been previously used to

explain approach and avoidance learning in the basal ganglia, which is driven by the dopamine modulation of Go and Nogo pathways, respectively. The tendencies to choose or inhibit actions are encoded in the strengths of synaptic connections between cortical neurons associated with the current context and the Go and Nogo neurons.

Learning of synaptic weights Positive and negative reward prediction errors facilitates Go and Nogo learning, respectively. In the OpAL model, the synaptic weights of Go and Nogo neurons were updated for each stimulus i using Eq. (2.10) and Eq. (2.11), respectively. On each trial, the model selected the action with the highest probability based on the expected positive and negative consequences of an action encoded in the G and N weights, respectively. For example, the probability of choosing A on an AB trial was computed as follows:

$$P_A = \frac{1}{1 + \exp(-(DV_A - DV_B))}, \tag{5.1}$$

where the decision value DV in the OpAL model of stimulus i is $DV_i = \beta_G G_i - \beta_N N_i$. The parameters β_G and β_N controlled the relative contribution of the Go and Nogo pathways, respectively. When $\beta_G = \beta_N$ both pathways contributed equally, when $\beta_G > \beta_N$ the Go pathway dominated, and when $\beta_G < \beta_N$ the Nogo pathway dominated.

The synaptic weights for G and N in the payoff-cost model were updated using Eq. (2.12) and Eq. (2.13), respectively. In contrast to the simulations in chapter 3, only positive prediction errors contributed to the updating of Go weights, and only negative predictions errors contributed to the updating of Nogo weights. The decision value in the payoff-cost model is $DV_i = (\beta_G + \beta_N)Q_i - (\beta_N - \beta_G)S_i$, which depends on the mean reward ($Q_i = G_i - N_i$), and the spread in reward outcomes ($S_i = G_i + N_i$). This decision value was also entered into the softmax rule, Eq. (5.1).

The simulations followed the task structure as described in the methods. First, the synaptic weights were learned in the training phase and then the learned 'cached' synaptic weights were used to guide choices during the testing phase. Synaptic weights for Go and Nogo neurons were initialised at 0.1 and updated using learning rates for Go and Nogo neurons $\alpha_G = \alpha_N = 0.1$. In the OpAL model, the value of a state was also learned by the critic according to Eq. (2.2) with learning rate, $\alpha_c = 0.1$. During training, the decision parameters were set $\beta_G = \beta_N = 2$, because I assumed that both pathways contributed equally. The model received a payoff of $r \in \{-1, 0, 1\}$, depending on the reward probability associated with a stimulus.

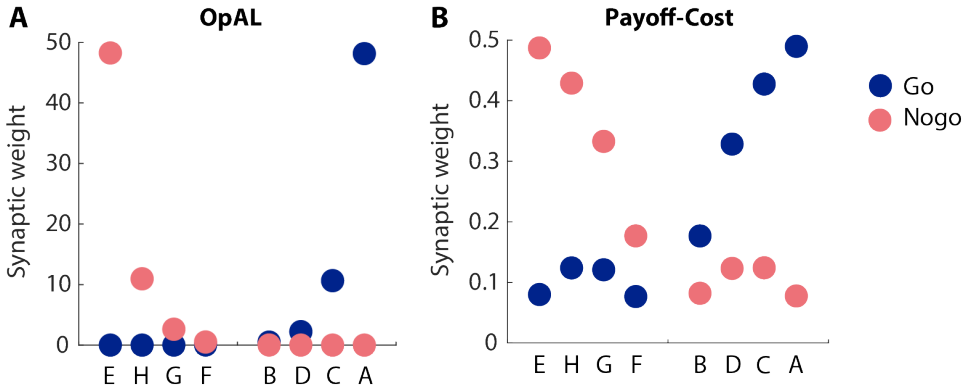


Figure 5.2: **Learned synaptic weights for Go and Nogo neurons.** A) In the Opponent Actor Learning (OpAL) model, synaptic weights for Go and Nogo neurons were learned using Eq. (2.10) and Eq. (2.11), respectively. B) In the payoff-cost model, synaptic weights for G and N were updated using Eq. (2.12) and Eq. (2.13), respectively. The decay parameter was set to $\lambda = 0.1$ and non-linearity parameter to $\epsilon = 0$. For both models, parameters were set to $\alpha_G = \alpha_N = \alpha_c = 0.1$, $\beta_G = \beta_N = 2$. Go and Nogo weights were initialised at $G_i = N_i = 0.1$.

Although both models learned weights according to the valence and probability of outcomes associated with a stimulus, they learned different values. In the OpAL model, G_i increased, while N_i decreased, with increasing reward probability. The relationship between synaptic weights and reward probabilities was non-linear, which resulted from the multiplication of the reward prediction error by G_i or N_i (Eq. (2.10) and Eq. (2.11)). The synaptic weights grew exponentially as the weights were multiplied by increasingly higher values. In

the payoff-cost model, G_i also increased, while N_i decreased, with increasing reward probability. However, this model also depended on the average reward of a stimulus ($Q_i = G_i - N_i$), and the spread in reward outcomes ($S_i = G_i + N_i$). Stimuli in 60/40 pairs have higher outcome variance compared to stimuli in 80/20 pairs, resulting in overweighting of positive and negative consequences of actions. Therefore, the relationship between synaptic weights and reward probabilities in the payoff-cost model is concave rather than convex as in the OpAL model.

Simulated test performance I assumed that the test performance depended on the learned synaptic weights during training. The tendency to choose an action was encoded in the weights of Go neurons, whereas the tendency to avoid an action was encoded in the weight of Nogo neurons. Additionally, choice behaviour during the test depended on the level of dopamine. Hunger increased dopamine levels and was simulated by enhancing the Go pathway in choices ($\beta_G > \beta_N$). Satiety reduced dopamine levels and was simulated by increasing the activity in the Nogo pathway ($\beta_G < \beta_N$). Given that the G and N weights in the OpAL model converged to values that were multiple orders of magnitude larger than those in the payoff-cost model, I used different values of β_G and β_N with the same ratio. For the OpAL model, I set $\beta_G = 0.2$, $\beta_N = 0.05$ for hunger and $\beta_G = 0.05$, $\beta_N = 0.2$ for satiety. For the payoff-cost model, I used $\beta_G = 4$, $\beta_N = 1$ for hunger and $\beta_G = 1$, $\beta_N = 4$ for satiety.

Note that the true difference in expected value between A and C is identical to that between D and B, and the difference between A and D is identical to that of C and B. Thus, the difference in performance between approach and avoidance behaviour reflects the bias for positive versus negative feedback. Standard reinforcement learning models, which converge to the theoretical expected value of a stimulus, produce equal Choose A and Avoid B performance, leading to zero bias.

In the OpAL and payoff-cost model, the accuracy for choosing A was mainly driven by the difference between A and C, $|G_A - G_C|$, and the accuracy for avoiding B was mainly driven by the difference between B and D, $|G_B - G_D|$. In the OpAL model, $|G_A - G_C| > |G_B - G_D|$ resulting in a higher accuracy for choosing A, whereas $|G_A - G_C| < |G_B - G_D|$ in the payoff-cost model, resulting in higher accuracy for avoiding B. Both models predict that high dopamine levels increase overall performance in the gain domain, because all Go weights are higher than Nogo weights $G > N$. In contrast, low dopamine levels increase overall performance for the loss domain, because all Nogo weights are higher than Go weights $G < N$. A positive bias results in approach behaviour and is linked to high levels of dopamine, and vice versa for a negative bias (Frank et al., 2007, 2004).

Theoretical predictions The introduction of a gain and loss domain provided additional predictions to the ones previously reported (Collins and Frank, 2014; Mikhael and Bogacz, 2016). In those simulations, correct (positive) and incorrect (negative) feedback were mapped onto a reward of 1 and -1, respectively. In my simulations, positive and negative feedback for the gain domain were represented by a reward of 1 and 0, respectively, whereas positive and negative feedback for the loss domain were represented by a reward of 0 and -1, respectively.

The simulations with the OpAL model and payoff-cost models revealed qualitative patterns. The OpAL model predicted a positive bias in the gain domain (Fig. 5.3A), but a negative bias in the loss domain (Fig. 5.3B). The direction of the bias was the same for both conditions, but the overall performance in the gain domain was higher for hungry, while the performance in the loss domain was higher for sated. The prediction for overall performance corresponds to previous reports (Fig. 5.3C), because this condition compared the synaptic weights learned for rewards with a value of 1 and -1. The payoff-cost model predicted a negative bias for hungry and positive bias for sated irrespective of

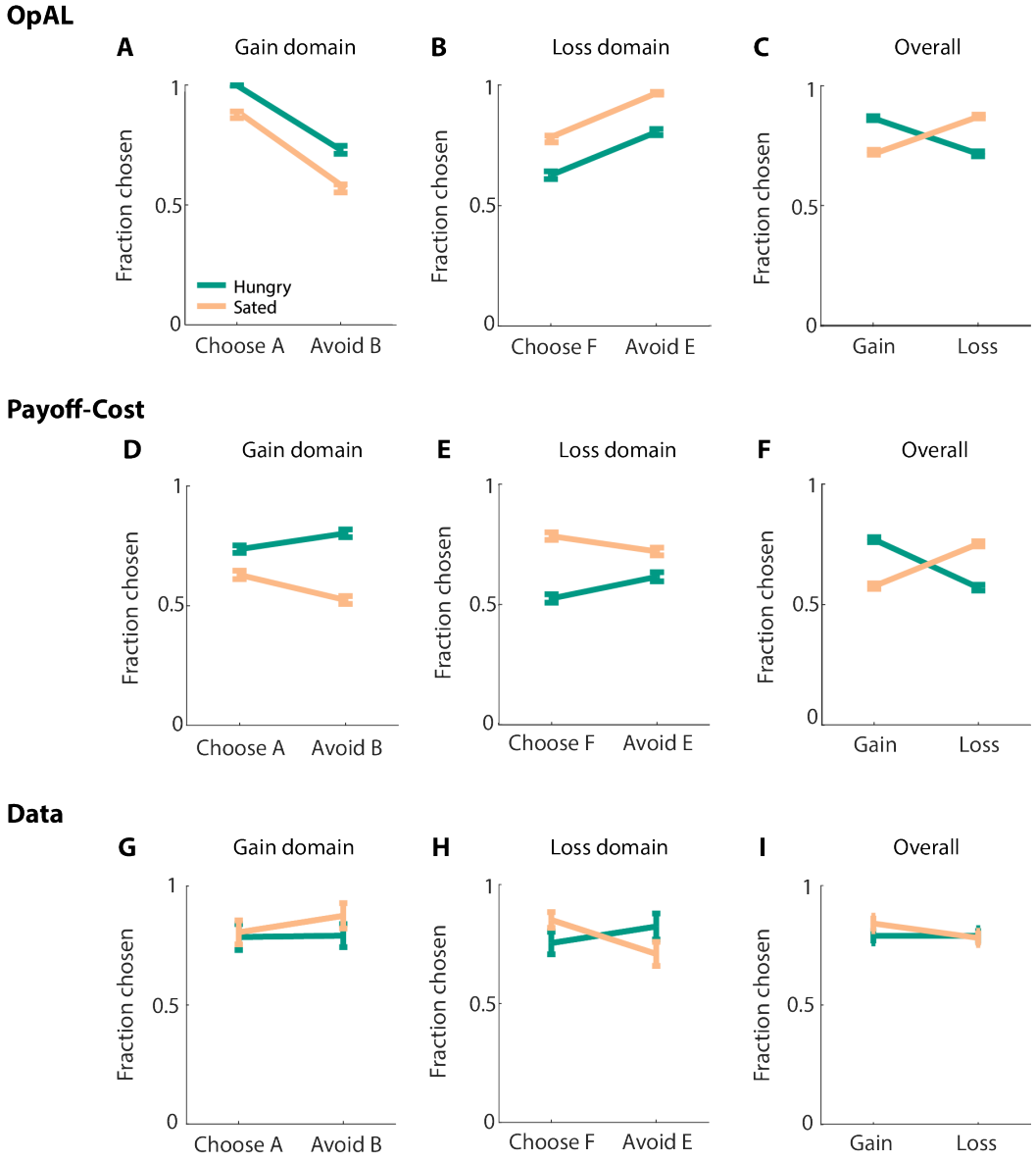


Figure 5.3: **Predicted and empirical approach-avoidance behaviour at test.** The total fraction chosen was calculated as the number of times the stimulus with the highest expected value was chosen or the stimulus with the lowest expected value was avoided. Overall performance was computed as the combined accuracy for both approach and avoidance behaviour in either the gain or loss domain. Predicted accuracy of 100 virtual participants by the OpAL model for (A) the gain domain (B) loss domain (C) combined performance. Predicted accuracy by the payoff-cost model for (D) gain domain, (E) loss domain, and (F) combined performance. The experimental data for (G) the gain domain, (H) loss domain, and (I) combined performance. Error bars represent SEM.

whether the choice was made in the gain or loss domain (Fig. 5.3D–E), which is consistent with previous studies (Mikhael and Bogacz, 2016). The overall performance for the gain and loss domain was the same as for the OpAL model (Fig. 5.3F).

The modification to the task design allowed me to test the effects of food deprivation on learning from positive and negative feedback in two ways; based on avoidance or approach behaviour and based on performance for positive or negative valence.

Food deprivation attenuated avoidance behaviour

Approach and avoidance behaviour I examined the effects of food deprivation on learning from positive and negative reward prediction errors using the analysis described by Frank et al. (2004). I used a three-way repeated measures ANOVA to examine which of the following factors, approach-avoidance behaviour, food deprivation or valence (gain/loss domain), affected test performance. Food deprivation modulated approach and avoidance behaviour differently (interaction effect of food deprivation and approach-avoidance [$F_{1,20} = 4.48$, $p = 0.047$]; Fig. 5.3G–H). Sated participants tended to perform better in the gain domain compared to the loss domain and compared to food deprived individuals (interaction effect of food deprivation and valence [$F_{1,20} = 3.01$, $p = 0.098$]; Fig. 5.3I). The main effect of food deprivation ($F_{1,20} < 1$), valence ($F_{1,20} = 2.91$, $p = 0.104$), approach-avoidance behaviour ($F_{1,20} = 1.91$, $p = 0.183$), or other interaction effects ($F_{1,20} < 1$) were not significant.

Given the significant interaction effect of food deprivation and avoid-approach behaviour, I further examined this effects using two-way repeated measures ANOVAs. For avoidance behaviour, participants performed differently for gains compared to losses (main effect of valence [$F_{1,20} = 4.45$, $p = 0.048$]). In comparison to the deprived condition, sated participants were better at avoiding infrequent gains compared to avoiding frequent losses (interaction effect of food deprivation and valence [$F_{1,20} = 7.867$, $p = 0.011$]). Food deprivation did not

alter overall avoidance behaviour (main effect of food deprivation [$F_{1,20} < 1$]). For approach behaviour, participants performed equally well regardless of valence or deprivation level [all $p > 0.2$].

Slowing Slowing is another type of avoidance behaviour, which is measured as the increase in reaction time for avoiding the worst stimulus (B and E) compared to choosing the best stimulus (A or F). Reaction times were only assessed on correct trials. Food deprived participants tended to show less slowing behaviour (main effect of food deprivation [$F_{1,20} = 3.92$, $p = 0.062$]; Fig. 5.4A), which was not affected by the valence of the outcomes (interaction effect of food deprivation and valence [$F_{1,20} = 1.46$, $p = 0.242$]). Slowing behaviour was not different for the gain and loss domain (main effect of valence [$F_{1,20} < 1$]). Interestingly, food deprivation increased overall reaction times on novel test pairs ([$t_{20} = 2.85$, $p = 0.010$]; Fig. 5.4B), but this did not affect the number of missed trials [$t_{20} = 1.37$, $p = 0.096$].

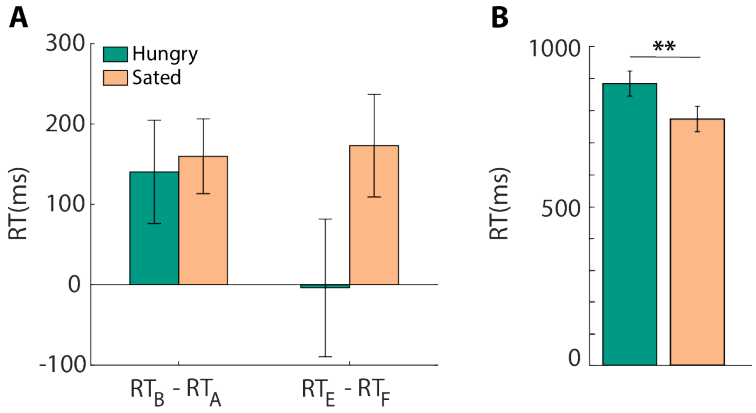


Figure 5.4: **Slowing and reaction times.** A) Slowing behaviour indicates an increase in reaction times for avoiding the least rewarding stimulus (B and E) compared to choosing the most rewarding stimulus (A or F). B) Overall reaction times for novel test pairs. Error bars represent SEM. ** $p < 0.01$.

Effect of difficulty In addition to the initial analyses, the novel testing pairs were further divided into two levels of difficulty, or conflict. High conflict pairs are novel pairs with two shapes that were most rewarding (or punishing) in the

original pairs (i.e. AC and FH as approach pairs and BD and EG as avoidance pairs). Low conflict trials are pairs in which only one of the shapes was the most rewarding stimulus in the original pairs (i.e. AD and FG for approach pairs and BC and EH for avoidance pairs). The difference in expected values of stimuli in high conflict pairs is small, which makes them more difficult than low conflict pairs that have larger differences in expected values. A small difference in expected values increases reaction times (Fontanesi et al., 2019), and is indicative of a higher difficulty level.

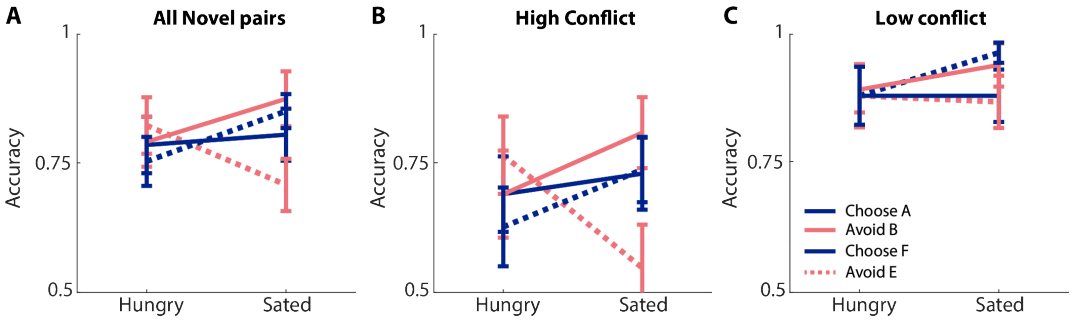


Figure 5.5: **Performance depended on the level of difficulty.** A) Avoidance and approach behaviour for the gain and loss domain in all novel pairs. B) Both stimuli in high conflict trials were the most rewarding in the original pairs, which makes the choice more difficult. Overall accuracy decreased and the spread in performance between the food deprived and sated condition was amplified. C) In low conflict pairs, only one stimulus was most rewarding in the original pairs, the difference in expected value was larger and the overall test performance on these pairs was higher. Differences in performance caused by food deprivation were smaller in low conflict pairs than in high conflict pairs. Error bars represent SEM.

To examine if conflict level altered test performance, I used a four-way repeated measures ANOVA with the following factors, conflict, approach-avoidance behaviour, food deprivation or valence (gain/loss domain). Participants were significantly better at low conflict trials compared to high conflict trials (main effect of conflict [$F_{1,20} = 35.87$, $p < 0.001$]), confirming the impact of difficulty on test performance. Test performance for approach and avoidance behaviour tended to be differently affected by valence and deprivation level (interaction effect of food deprivation $\times$ valence $\times$ approach-avoidance [$F_{1,20} = 3.93$, $p = 0.061$]). All other main or interaction effects were not significant [$F_{1,20} < 1$].

Food deprivation did not attenuate learned values

As discussed above, test performance could be driven by changes in the learned values or by differences in weighting of learned values for choice behaviour during the test. To verify whether the effect of food deprivation on test performance was the result of altered learning, I examined whether the accuracy of learned outcome probabilities was affected by the level of food deprivation. The proportion of correct trials was counted as the number of times the most rewarding (or least punishing) stimulus was chosen in each of the pairs for both conditions.

Learned values During the training phase, participants received feedback for both the factual and the counterfactual stimulus, which typically improves performance (Palminteri et al., 2015). A full feedback paradigm has the additional consequence that it is not required to sample options for feedback to behave optimally. Learning can occur independently of choice behaviour, which means that the optimal strategy is to exploit the best option and not explore the alternative option. This is in contrast to partial feedback design, where exploration is necessary to learn the reward distributions of the presented stimuli and optimise behaviour.

All participants performed an equal number of trials in both sessions, which is different from the original design (Frank et al., 2004), but consistent with imaging studies (e.g. Jocham et al. (2011); Shiner et al. (2012)). The number of missed trials was not significantly different between conditions [$t_{20} = 1.70$, $p = 0.104$], suggesting that participants received a similar amount of feedback to learn the meaning of the stimuli.

On average, participants learned to perform more accurately over time and they learned faster on the easier AB/EF trials than the more difficult CD/GH trials (Fig. 5.6A–D). Performance stabilised in the last block (last 20 trials for each pair). A three-way repeated measures ANOVA revealed

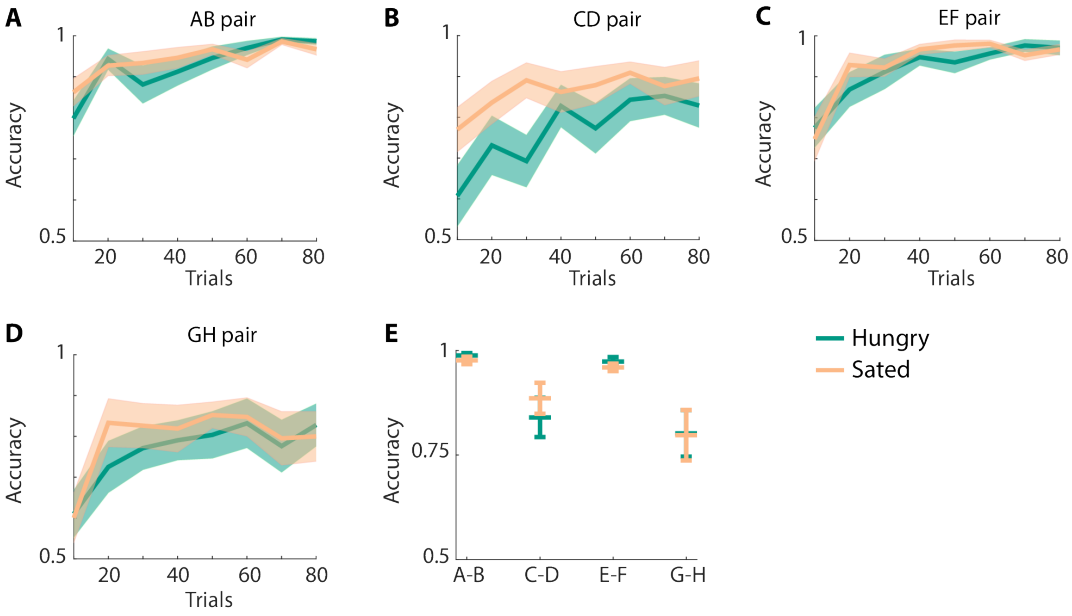


Figure 5.6: **Binned proportion of correct trials during training.** Accuracy during training was calculated as the overall number of correct choices for the most rewarding or least punishing stimulus in bins of 10 trials. Performance was above chance level, indicating that people reliably chose the correct stimulus on the majority of the trials. A) Learning for AB pair. B) Learning for CD pair. C) Learning for EF pair. D) Learning for GH pair. E) Group average accuracy for each pair calculated for the last 20 trials only. Error bars represent SEM.

that the overall accuracy in the last block was comparable for both sated and food deprived individuals (main effect of food deprivation [$F_{1,20} < 1$]; Fig. 5.6E). Independently of the deprivation level, participants were overall better at learning the reward contingencies in the gain domain (main effect of valence [$F_{1,20} = 5.99$, $p = 0.024$]) and for the easier 80/20 stimuli (main effect fo difficulty [$F_{1,20} = 29.86$, $p < 0.0001$]). Food deprivation did not affect difficulty [$F_{1,20} = 2.64$, $p = 0.120$], valence [$F_{1,20} < 1$] or a combination of valence and difficulty [$F_{1,20} < 1$]. In summary, valence and difficulty modulated learned values, but food deprivation did not.

Stay-switch behaviour The differences in how values were learned during training may be driven by an effect of food deprivation on learning rate. I used a stay-switch metric to examine this possibility. A stay-switch metric provides insight into whether food deprivation altered the tendency to stay or shift after

receiving positive or negative feedback. I counted how often the participant chose the same stimulus again after receiving positive (win 0.50 for AB and CD and lose 0 for EF and GH) or negative (win 0 for AB and CD and lose 0.50 for EF and GH) feedback for the chosen option the next time this pair was presented. The proportion of stay behaviour was calculated for each pair and divided by the total number of times the feedback was positive (or negative) to give the proportion for each condition.

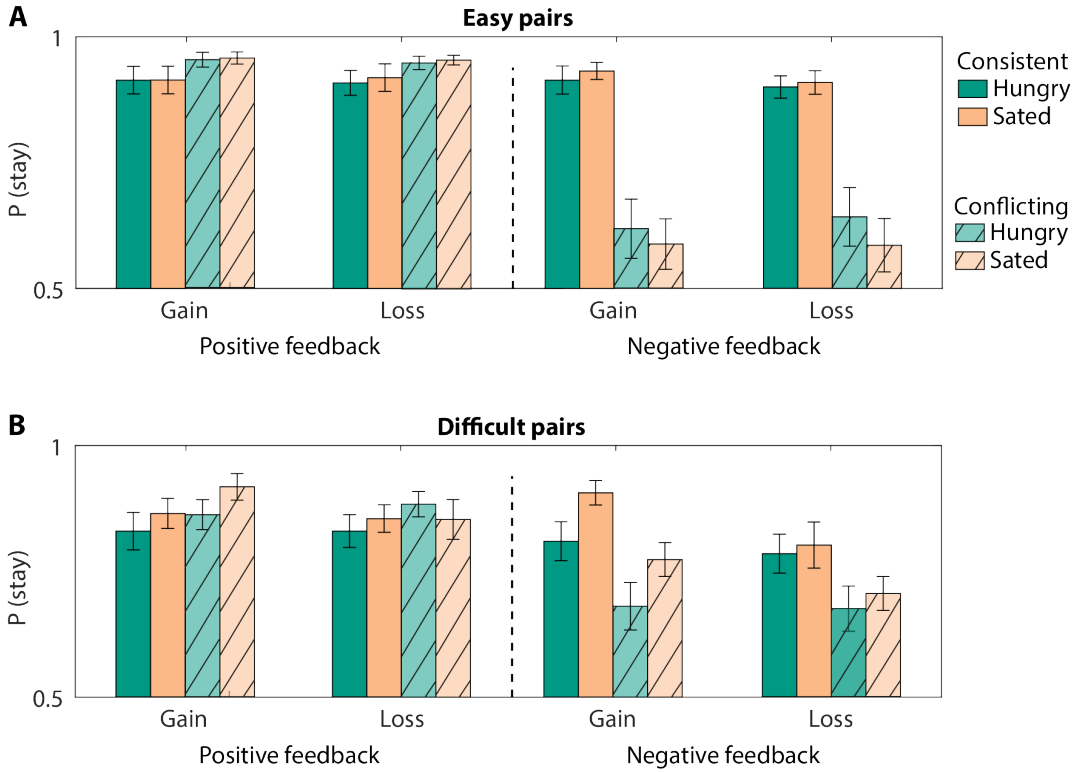


Figure 5.7: **Stay probabilities after receiving positive or negative feedback on previous trial.** Stay behaviour for the gain and loss domain after receiving positive or negative feedback on the previous trial the pair was presented. Consistent feedback indicates that feedback for both stimuli was the same. Conflicting feedback indicates that the feedback for the chosen and unchosen stimulus were different. Easy pairs were AB and EF and difficult pairs were CD and GH. Stay probabilities for easy and difficult pairs are depicted in panel A and B, respectively. Error bars represent SEM.

The presence of both factual and counterfactual feedback could influence stay behaviour, particularly when the feedback of the two stimuli is different. If factual and counterfactual feedback is the same, the outcome on the cur-

rent trial will have minimal impact on stay behaviour on the following trial. If the factual feedback is positive, but the counterfactual feedback is negative, the feedback reinforces the current choice and makes individuals more likely to stay (confirmatory feedback). If the factual feedback is negative, and the counterfactual feedback is positive, the feedback is punishing and increases the likelihood of switching on the next trial (disconfirmatory feedback). This type of behaviour is commonly observed in full feedback tasks and creates a confirmation bias (Palminteri et al., 2017). The 80/20 pairs include more confirmatory feedback that reinforce choosing the most optimal stimulus, whereas the 60/40 pairs are more uncertain, because the disconfirmatory evidence is higher.

The consistency of the feedback of the presented stimuli indeed modulated stay behaviour for easy and difficult pairs (Fig. 5.7). Stay behaviour was also modulated by other factors, including the difficulty of the pair, whether the feedback on the previous trial was positive or negative and the deprivation level. However, the valence of the domain did not affect stay behaviour. In the next section, I employ computational modelling techniques to examine whether any of these factors affected learning and were modulated by food deprivation.

Computational modelling The accuracy and stay-switch behaviour during the training phase suggest that food deprivation may influence how people choose and learn about the outcomes of stochastic stimuli. However, these one-trial-back analyses do not provide insight into the computational processes involved in the integration over a longer reward history. This is especially true for a full feedback design, in which learning could occur regardless of which option is chosen. Therefore, I used reinforcement learning models to elucidate what drove the behavioural changes observed in the data.

I used four reinforcement learning models, which are typically used for full feedback paradigms (Palminteri et al., 2015), to test and confirm whether food deprivation altered the learning rate or made participants more random in their

	Positive feedback	**Negative feedback**
Full		
Factual	$Q_c = Q_c + \alpha_c^+ \delta_c$	$Q_c = Q_c + \alpha_c^- \delta_c$
Counterfactual	$Q_u = Q_u + \alpha_u^+ \delta_u$	$Q_u = Q_u + \alpha_u^- \delta_u$
Information		
Factual	$Q_c = Q_c + \alpha_c \delta_c$	$Q_c = Q_c + \alpha_c \delta_c$
Counterfactual	$Q_u = Q_u + \alpha_u \delta_u$	$Q_u = Q_u + \alpha_u \delta_u$
Valence		
Factual	$Q_c = Q_c + \alpha_+ \delta_c$	$Q_c = Q_c + \alpha_- \delta_c$
Counterfactual	$Q_u = Q_u + \alpha_+ \delta_u$	$Q_u = Q_u + \alpha_- \delta_u$
Confirmatory		
Factual	$Q_c = Q_c + \alpha_{\text{con}} \delta_c$	$Q_c = Q_c + \alpha_{\text{dis}} \delta_c$
Counterfactual	$Q_u = Q_u + \alpha_{\text{dis}} \delta_u$	$Q_u = Q_u + \alpha_{\text{con}} \delta_u$

Table 5.1: **Update rules for computational models.** Q-values for the chosen (Q_c) and unchosen (Q_u) were updated according to the equations listed in the table. The reward prediction error for chosen stimulus was $\delta_c = r_c - Q_c$ and unchosen stimulus was $\delta_u = r_u - Q_u$, where r_c and r_u are the outcomes for the chosen and unchosen stimulus respectively. Choice probabilities were computed using the softmax decision rule, Eq. (2.14).

choice behaviour. These models rely on standard Q-learning and the softmax choice rule, Eq. (2.14). The models describe how positive and negative feedback for factual and counterfactual stimuli are incorporated to learn the average reward of each of the stimuli. The learning rules for the Q-values of the 4 models are outlined in Table 5.1. The full model consists of 5 parameters: one softmax parameter and 4 learning rules for positive and negative feedback for both factual and counterfactual stimuli. The information model is a reduced version of the full model with only 3 parameters: one softmax parameter, a learning rate for factual feedback, and a learning rate for counterfactual feedback. This model tests whether individuals focus on obtained or foregone outcomes. The valence model also has 3 parameters: one softmax parameter, a learning rate for positive feedback, and a learning rate for negative feedback. This model tests whether individuals learn more from positive or negative outcomes. The confirmatory model also has 3 parameters: one softmax parameter, one learning rate for confirmatory feedback, and one for disconfirmatory feedback. This

model tests whether positive or negative feedback affects factual and counterfactual learning in opposing directions, such that negative unchosen prediction errors are more likely to be taken into account than positive unchosen prediction errors. In other words, it tests if there is a bias toward information that supports the current choice. The behavioural data (Fig. 5.7) suggest that this model could be a good candidate for explaining behavioural data during learning. Lastly, I also fitted the two models grounded in the basal ganglia architecture that I used to simulate this task at the beginning of this chapter.

I used a hierarchical model fitting procedure as described in the Methods (section 4.3). The model parameters, denoted by Greek letters, were transformed onto a Gaussian scale, denoted by their respective Latin letters, using Eq. (4.7) for the learning rates and Eq. (4.9) for the softmax parameter. The statistical significance was tested with respect to the Gaussian scaled model parameters. Model comparison (see Methods section 4.3) of the six models showed that the behavioural data during training was best explained by the confirmatory model (Fig. 5.8A). The quality of the fitting procedure was verified with a parameter recovery analysis (see Methods section 4.3). All parameters were well recovered ($R > 0.8$) and the model fitting procedure did not introduce spurious correlations between the other parameters ($|R| < 0.32$; Fig. 5.8B).

The confirmatory learning rate was significantly higher than the disconfirmatory learning rate for all participants, regardless of deprivation level (main effect of learning rate [$F_{1,20} = 440.68$, $p < 0.0001$]; Fig. 5.8C). This suggests that participants have the tendency to learn faster from outcomes that support their current choice, which is consistent with previous reports on human choice behaviour in a full feedback design (Palminteri et al., 2017). Food deprivation increased overall learning (main effect of food deprivation [$F_{1,20} = 8.132$, $p = 0.010$]). Pairwise comparison of the parameter estimates revealed that food deprivation increased the confirmatory learning rate [$t_{20} = 2.33$, $p = 0.030$] and caused a trend increase for the disconfirmatory learning rate [$t_{20} = 2.05$,

$p = 0.054$]. However, the significant effect for confirmatory learning rates disappeared after Bonferroni adjustment. Food deprivation did not alter the softmax parameter ([$t_{20} = -0.99$, $p = 0.334$]; Fig. 5.8D), suggesting that the level of food deprivation did not alter the tendency to explore or exploit options. This also supports the absence of strong significant effects in the stay-switch analyses.

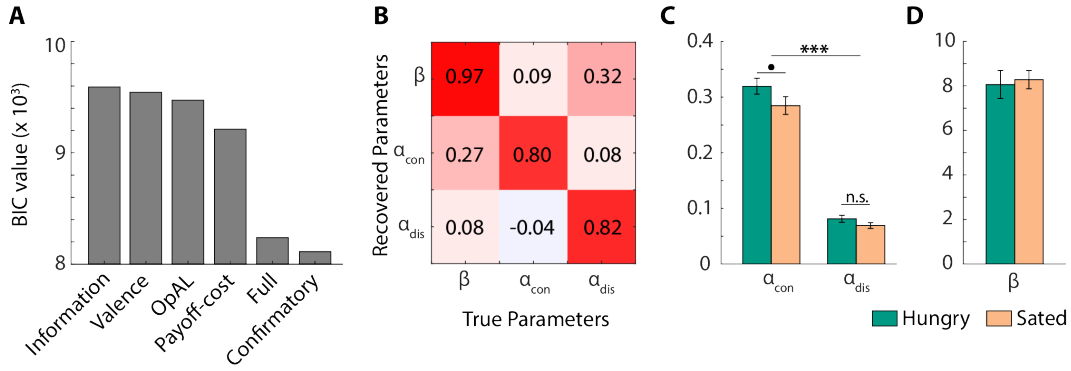


Figure 5.8: **Computational modelling results.** A) BIC values for the models fitted to the experimental data. Best fitting model was the confirmatory model, indicated by the lowest BIC value. B) Correlation matrix of the free parameters used to generate the simulated data (True parameters) and the obtained parameters by applying the parameter estimation procedure on the simulated data (Recovered parameters). Bright red values indicate a strong correlation between the true and recovered parameter value and therefore a good parameter recovery. C) Learning rates for confirmatory feedback and disconfirmatory feedback, α_{con} and α_{dis}, respectively. D) Softmax temperature, β, derived from confirmatory model. Statistical significance was tested on Gaussian scaled parameters. $\bullet < 0.05$, $*** < 0.001$. Error bars represent SEM.

Using the best-fitting parameters of the confirmatory model, I estimated the latent Q-variables of the different stimuli for both sessions. The learned Q-values at the end of the reinforcement learning phase showed a decrease with decreasing reinforcement probability (main effect of symbol [$F_{7,140} = 1355.46$, $p < 0.0001$]; Fig. 5.9). Thus, the best option had also acquired the highest Q-value and all values corresponded to their respective objective expected values. Food deprivation did not alter the learned Q-value estimates (main effect of food deprivation [$F_{1,20} = 1.95$, $p = 0.178$] or interaction effect of food deprivation and stimulus [$F_{7,140} < 1$]), suggesting that the effects seen in the test did

not arise from differences in learned values during training. This is in line with Fig. 5.7E, which shows that there was no difference in overall accuracy in the last block of training between the food deprived and sated condition.

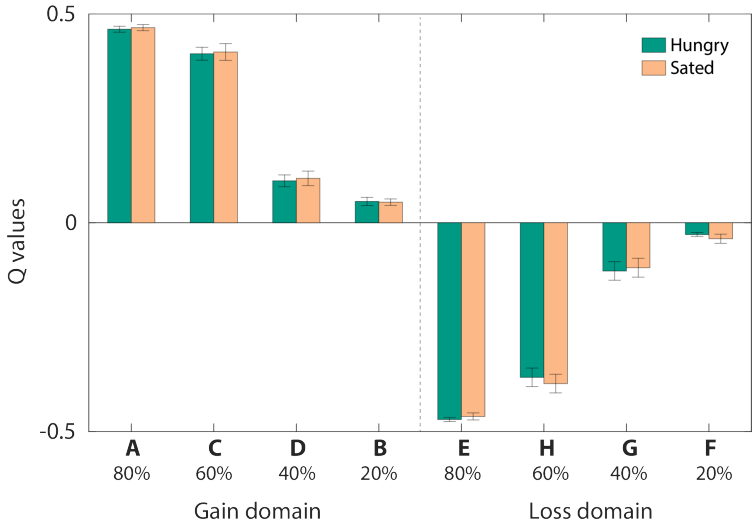


Figure 5.9: **Food deprivation did not alter learned Q-values.** Q-values were generated using the parameter estimates obtained from the confirmatory model. Q-value estimates corresponded to their respective objective expected values. Food deprivation did not alter the learned Q-value estimates. Error bars represent SEM.

5.4 Discussion

In this study, I investigated whether food deprivation altered approach and avoidance behaviour. Based on the theory described in chapter 3, I predicted that food deprivation enhances learning from positive feedback and satiety promotes learning from negative feedback. I found that food deprivation differentially modulated approach and avoidance behaviour, but not in the direction predicted by my theory (Fig. 5.3). Food deprived participants performed equally well for avoiding or approaching outcomes with positive or negative valence, whereas sated individuals were better at avoiding infrequent gains compared to avoiding frequent losses. The learned stimulus values during training were affected by valence and outcome uncertainty, but not by food deprivation. Computational modelling confirmed that the acquired values at the end

of training were unaffected, despite the increased learning rate of hungry individuals and the small differences in sensitivity to positive and negative feedback. This study showed that food deprivation altered action selection, without affecting learning. The opposing effects of food deprivation on avoidance behaviour for positive and negative outcomes suggests this behaviour may not be solely driven by the function of dopamine in the basal ganglia as suggested in chapter 3.

Comparison of computational models

Test performance in the probabilistic selection task has been described with at least three different types of models, the OpAL model, payoff-cost model and a standard Q-learning model. I address the similarities and differences of these models below.

The theoretical interpretation of the OpAL model is as follows. Each time a positive (good) outcome is experienced, the synaptic weight of Go neurons increase. Each time a negative (bad) outcome is experienced, the synaptic weight of Nogo neurons increase. These synaptic weights grow exponentially and tend to infinity with significant training. Consequently, the stimulus with the highest Go weight is chosen and the stimulus with the highest Nogo weight is avoided. The learning rate parameter controls the growth of a synaptic weight on each trial, meaning that high learning rates lead to faster updating (learning) and higher synaptic weights. This suggests that high learning rates lead to high synaptic weights and drive either approach or avoidance behaviour. A drawback of this model is that exponentially growing weights reach values that are difficult to interpret biologically. The predicted patterns in this chapter are different to patterns previously reported (Collins and Frank, 2014; Mikhael and Bogacz, 2016), because I considered outcomes with positive, negative and zero value. A reward value of zero causes increasingly smaller reward prediction errors. In the update rules for synaptic weights, the current weight is multiplied by the reward prediction error, until Go weights in the loss domain and Nogo

weights in the gain domain converge to zero. Dopamine controls the overall accuracy, but does not change the direction of the bias within a domain, because dopamine only scales non-zero weights.

The theoretical interpretation of the payoff-cost model differs slightly from the OpAL model. Go and Nogo neurons also learn based on the frequency of positive and negative outcomes. However, these neurons converge to values that are proportional to the expected value of a stimulus, which is calculated as the reward probability multiplied by the payoff. In addition, higher uncertainty results in overweighting of positive and negative consequences of actions. This interpretation fits with behavioural data on risk-taking behaviour where both the expected value and uncertainty in outcomes drive choices. The difference in Go and Nogo weights of the available stimuli drives approach and avoidance behaviour, respectively. High dopamine promotes avoidance behaviour, while low dopamine enhances approach behaviour, independently of whether the outcomes had a positive or negative valence.

The standard Q-learning model is not grounded in the basal ganglia architecture. Each stimulus has one single value that reflects the integrated history of both positive and negative contingencies. The expected values of a stimulus is learned, without considering the uncertainty in outcomes. Choices are made based on the relative differences in Q-values of the available stimuli. These models do not predict any differences between approach and avoidance behaviour, because the difference between the expected value of A and C or D is equal to the difference in expected values between of B and D or C. This logic applies to both the gain and loss domain.

Regardless of learning rate and with sufficient training, the payoff-cost and Q-learning models eventually converge to values that are proportional to the expected value of a stimulus. In contrast, the converged values of the synaptic weights in the OpAL model depend on the size of the learning rate and the number of training trials. Although the OpAL and payoff-cost model provide

mechanistic insights, the behavioural data is better explained by simpler Q-learning models.

Effect on learning

To examine the effect of food deprivation on learning, I used the OpAL model, payoff-cost model, and four variants of a Q-learning model. The valence model, which is the most used Q-learning model for this task, did not explain the training data well. Changing the partial feedback design to a full feedback design may have contributed to this observation. However, I can still draw parallels between the two designs and compare how individuals learn from reinforcing and punishing feedback. In the full feedback design, participants considered both factual and counterfactual feedback, which was best captured by the confirmatory model in the current study. The feedback was reinforcing when the factual feedback was positive and the counterfactual feedback was negative. In contrast, the feedback was punishing when the factual feedback was negative and the counterfactual feedback was positive. Similar to the valence model, reinforcing feedback promotes choice repetition and punishing feedback promotes choice switching.

Dopamine is thought to modulate Go and Nogo neurons in the basal ganglia to facilitate and suppress the execution of cortical actions, respectively. Go signals may trigger the updating of working memory representations in the prefrontal cortex, whereas Nogo signals prevent updating to maintain previously stored information (Frank, 2005). Positive reinforcements drive dopamine bursting that enhance Go firing and learning, whereas negative reinforcements have the opposite effects. Food deprived individuals have higher learning rates, suggesting that they are more sensitive to recent outcomes. This effect tended to be higher for reinforcing, confirmatory, feedback compared to negative, disconfirmatory feedback, suggesting that food deprivation increases learning from positive feedback and enhances the bias for information that supports the current choice. Although this effect is not very strong, it would fit with the

hypothesis that food deprivation promotes learning from positive feedback by enhancing the dopaminergic signal. Given that the performance during the training phase is thought to be mediated by a cooperative mechanism involving both the basal ganglia and the prefrontal cortex, it would be interesting to explore the links between food deprivation and working memory function in more detail.

In the theory, I assumed that food deprivation only affects the valuation (perceived utility) and learning of relevant reinforcers, which is in line with the valuation results presented in Fig. 4.2. Previous studies have suggested that motivational drives, such as hunger, could also affect decision-making for irrelevant monetary outcomes (Briers et al., 2006; de Ridder et al., 2014; Levy et al., 2013; Simon et al., 2017; Symmonds et al., 2010). Although these studies do not consider learning, the current study suggests that metabolic signals may affect learning for money as well, possibly by increasing dopaminergic signalling in the brain.

Effect on action selection

Choice behaviour during the test is more difficult to explain with computational models. The number of trials in the test is limited, which makes the interpretation and parameter estimation more challenging. Some studies employed two parallel Q-learning systems to fit the training data and test performance separately (Aberg et al., 2016; Frank et al., 2007). These two systems represent the computation executed by the basal ganglia and prefrontal cortex system, which have been shown to cooperate and compete for behavioural control (Daw et al., 2005; Frank and O'Reilly, 2006; Ostlund and Balleine, 2005). As discussed in the previous section, the prefrontal cortex is thought to guide rapid behavioural adjustments during training. The basal ganglia system may be better suited to differentiate between subtle differences in reinforcement values during the test as participants were told to rely on their "gut-instinct" when

making choices (Cavanagh et al., 2011; Frank et al., 2007). The theory described in chapter 3, only considered the contribution of the basal ganglia to action selection in this task, and does not make clear predictions about the involvement of the prefrontal cortex system. Food deprivation has been suggested to exert motivational control via the mPFC (Moscarello et al., 2007, 2010), which may contribute to the increased learning rates during training. However, this possibility was not addressed in chapter 3.

Despite not employing such modelling strategies, I observed that food deprivation affected biases for approach or avoidance behaviour for positive or negative valence outcomes. Sated individuals exhibited a bias for avoidance behaviour in the gain domain, but a bias for approach behaviour in the loss domain. High conflict choices amplified these biases, whereas low conflict choices decreased these biases. Hungry individuals did not show substantial biases for valence or approach-avoidance behaviour. The observed test data did not correspond clearly with the biases predicted by the models. These data suggest that the context decisions are made in, i.e. gains versus losses, may play an important role in approach and avoidance behaviour.

Neural substrates

The basal ganglia is one of the key areas that underlie behaviour in this task. Recent empirical data have shown that approach behaviour is linearly correlated with D1-receptor availability. In contrast, the relationship between avoidance behaviour and the binding of dopamine to the D2-receptor in the Nogo pathway follows an inverted-U shape (Cox et al., 2015). These data highlight that the theoretical predictions and assumptions may be an oversimplification of a complex system. Individual differences in tonic dopamine levels and asymmetric binding to D2-receptors in the putamen and frontal cortex predict motivational biases (Tomer et al., 2008, 2014), which is unaccounted for in the computational models. Higher D2-receptor binding in the left hemisphere is associated with preference for rewarding events, while a relatively

higher binding in the right hemisphere predicts a stronger tendency to avoid aversive outcomes (Tomer et al., 2008, 2014).

The D2-receptor activity has also been linked to many neurological disorders and seems to play an important role in avoidance behaviour. In a feeding related context, obese individuals have a lower availability of D2-receptors (Volkow et al., 2011). Food deprivation increases the activity of D2-receptors in a rodent model of obesity (Thanos et al., 2008), suggesting that the effects observed in this task may also be influenced by the total body mass of an individual. This will need to be investigated in future studies that include participants from a wider range of body weights.

An alternative explanation for the performance during the test and the effect of food deprivation on avoidance behaviour may extend beyond the basal ganglia circuitry. The vast majority of studies using probabilistic rewards have focused on decisions in which all possible outcomes are non-negative (McClure et al., 2004; Padoa-Schioppa and Assad, 2006; Preuschoff and Bossaerts, 2007; Tobler et al., 2007). Few studies have examined how decisions are made when all outcomes are in the domain of losses. The utility of positive and negative outcomes are thought to be evaluated by two systems in the brain. The VS and OFC are associated with the exclusive evaluation of gains (Mirenowicz and Schultz, 1996), whereas the amygdala and insula are associated exclusively with the evaluation of losses (Yacubian et al., 2007). Instead of relying on the basal ganglia system, avoidance behaviour may, in part, be driven by a circuitry that involves the amygdala. The BLA receives input from various feeding related brain areas. It plays an important role in hedonic valuation (Tiedemann et al., 2020) and satiety-dependent outcome devaluation (Parkes and Balleine, 2013). The BLA expresses ghrelin receptors (Huang et al., 2016) and sends feedback to sensory regions that are modulated by satiety (Burgess et al., 2016; Livneh et al., 2020). Increased homeostatic need resulting from acute food deprivation modifies microcircuits in the BLA to shift the balance

to positive valence encoding neurons, whereas satiety shifts the balance to negative valence encoding neurons (Calhoon et al., 2018). This suggests that the BLA may play an important role in the bias for approach and avoidance behaviour as a result of changes in metabolic need.

Finally, food-deprivation also acts as a mild stressor (Deroche et al., 1995), and has been shown to increase corticosterone levels in adolescent and adult rodents (Anderson et al., 2013). Stress has been linked to altered reward sensitivity (Berghorst et al., 2013; Lighthall et al., 2013), which may also contribute to the effects seen in this study.

Limitations

The interpretation of the current data set may be limited by the relatively small number of participants. The recruited number of participants was based on previous studies, but unfortunately more participants had to be excluded than anticipated due to chance level performance and incomplete datasets. Another limitation is that the introduction of a gain and loss domain has made the task more challenging than the original design. Despite this, the accuracy on the original pairs during the test suggests that the added difficulty did not alter the overall performance. Furthermore, the full feedback reduced the uncertainty in the task, making choices more deterministic. On one hand this could facilitate learning, but on the other hand, participants could form the wrong beliefs and choose the incorrect stimulus more often. Nevertheless, the correct average reward could still be learned. Computational modelling suggests that the average value can be learned even when accuracy during training is relatively low. Consequently, this could complicate the interpretation of the accuracy during training and comparison with the accuracy on test trials.

Conclusions

This study showed that the level of food deprivation affected approach and avoidance behaviour differently. The learned outcome probabilities were not

significantly different, even though food deprived participants had higher learning rates. Food deprivation affected how learned values were used to guide choice behaviour, but it is likely that these effects were not solely driven by the basal ganglia circuitry.

The reward is in the risk.

— Rachel Cone

6

The influence of metabolic state on risk-taking

Contents

6.1 Introduction 131

6.2 Methods 135

Experimental design 135

Behavioural analyses 138

Computational modelling 139

6.3 Results 139

Participants learned different reward distributions 139

Food deprivation altered learned risks, but not described risks 140

Modelling of risk-sensitive choice behaviour 142

Computational modelling captured risk preferences 146

Reaction times did not explain changes in risk attitudes . . 149

6.4 Discussion 150

Experience versus description 151

Effects of food deprivation 152

Neural processing of risky decision-making 155

Conclusion 158

6.1 Introduction

Daily changes in the environment have forced animals to adapt their behaviour to this variability when acquiring resources, particularly when resources are scarce. The strategies that guide humans and animals in their exploitation of variable environments are referred to as risk attitudes, where risk is defined as the uncertainty in the possible outcomes of a decision. The theoretical work described in chapter 3 did not explicitly outline the effects of food deprivation on learning and choice of uncertain outcomes. However, under the assumption that food deprivation increases dopamine levels, the theory also makes predictions about the effects of food deprivation on risk attitudes.

Dopamine is an important modulator of risk-taking behaviour. Dopamine-enhancing drugs increase risk-taking behaviour in rodents (Orsini et al., 2015; St Onge and Floresco, 2009), healthy individuals (Rigoli et al., 2016), and PD patients (Gallagher et al., 2007). Dopamine enhances reward seeking via the Go pathway, while decreasing the sensitivity to negative consequences encoded in the Nogo pathway (Frank et al., 2004). Individual differences in risk-seeking are linked to the expression of Nogo neurons (Zalocusky et al., 2016) and stimulation of these neurons can make previously risk-seeking rats risk-averse (Stopper et al., 2014). These observations suggest that food deprivation could increase risk-seeking through its control of dopaminergic signalling to increase the activity in the Go pathway, while reducing the activity in the Nogo pathway. In this chapter, I aim to answer the following question: Does food deprivation alter risk preferences for learned or described risks in various decision contexts?

Preferences in risk-taking result from cognitive biases that cannot be explained by rational behaviour. A rational decision-maker that tries to maximise expected reward would always choose the option with the highest utility regardless of the uncertainty in outcomes (von Neumann and Morgenstern, 1944). When choosing between two options with equal expected value that differ in risk, a rational decision-maker should prefer both options equally. Empirical

studies provide evidence that humans and animals show preferences that differ from those of a rational decision-maker. These observations have been used to develop new decision-making theories that incorporate attributes that influence the objective utility of a choice (similar to the utility function defined in chapter 3). Attributes, such as the past and present context of experiences, wealth and emotions or desires (e.g. current physiological and psychological status), define a reference point, or adaptation level, and choices are perceived in relation to this reference point (Helson, 1964).

The majority of empirical studies in animals showed that food deprivation increased risk-seeking (Kacelnik and Bateson, 1996), which led to the development of the risk-sensitive foraging theory (Houston, 1991; Kacelnik and Bateson, 1997; Stephens, 1981). This theory describes choice behaviour with respect to a metabolic reference point for options with equal expected value and different risks. Choosing the risky option should be preferred when energy levels are low, because this option increases the chance of exceeding the daily intake rate. When energy levels are high, the less variable option should be preferred, because this option decreases the chance of dropping below the daily intake rate.

In humans, risk patterns for economic decisions have been reported across a broad variety of contexts and do not always correspond to the risk patterns observed in animals. Animals need to learn about uncertain outcomes via trial-and-error, whereas humans can be exposed to risky decisions in an implicit or explicit manner. Some studies use abstract models of real life situations (e.g. lottery gambles) in which the probability and payoff magnitude associated with a gamble are explicitly described to the participant using words, numbers or diagrams (decisions from description) (Elliott et al., 1999; Hsu et al., 2005, 2009; Huettel et al., 2005, 2006; Kuhnen and Knutson, 2005; Preuschoff et al., 2006, 2008; Tom et al., 2007). This is in contrast to studies in which participants need to learn the different outcomes of choice alternatives by repeatedly

sampling them (decisions from experience) (Hertwig et al., 2004; Hertwig and Erev, 2009; Niv et al., 2012).

The typical risk patterns observed in humans may arise from the heuristic approach humans tend to use when integrating large amounts of information successfully (Gilovich et al., 2002). A heuristic approach for choice behaviour relies on the direct comparison of options, rather than the calculation of value. The comparison of a potential prospect to the reference point gives rise to positive and negative decision contexts. The biasing effect of decision contexts on choice behaviour are known as framing effects (Tversky and Kahneman, 1981). Framing effects for positive and negative decision contexts cause opposite and asymmetrical risk patterns (Kahneman and Tversky, 1979). These risk patterns are reversed for decisions from experience compared to decision from description (Hertwig and Erev, 2009), an observation that is known as the description-experience gap. In decisions from description, humans are generally risk-averse for decisions above a (monetary) reference point and risk-seeking for decisions below the reference point (Kahneman and Tversky, 1979). Prospect theory shows that low probability events are overweighted and high probability events are underweighted (Tversky and Kahneman, 1992). In decisions from experience, the reverse pattern has been found (Hertwig, 2012; Hertwig et al., 2004): people behave as if rare events have much *less* impact than they would objectively deserve. The two modes of decision-making have also been shown to depend on neural processes that are, at least partially, distinct (Fitzgerald et al., 2010; Jessup and O'Doherty, 2010).

Interestingly, a transfer from the metabolic domain onto the cognitive domain has been observed in human decision-making under uncertainty. Low energy levels, captured by an increased concentration of the hormone ghrelin, increased risk-seeking for food rewards *and* monetary outcomes (Levy et al., 2013; Shabat-Simon et al., 2018; Symmonds et al., 2010).

Risk attitudes in description-based and experience-based tasks are rarely compared in the same individual. Research on the effects of food deprivation

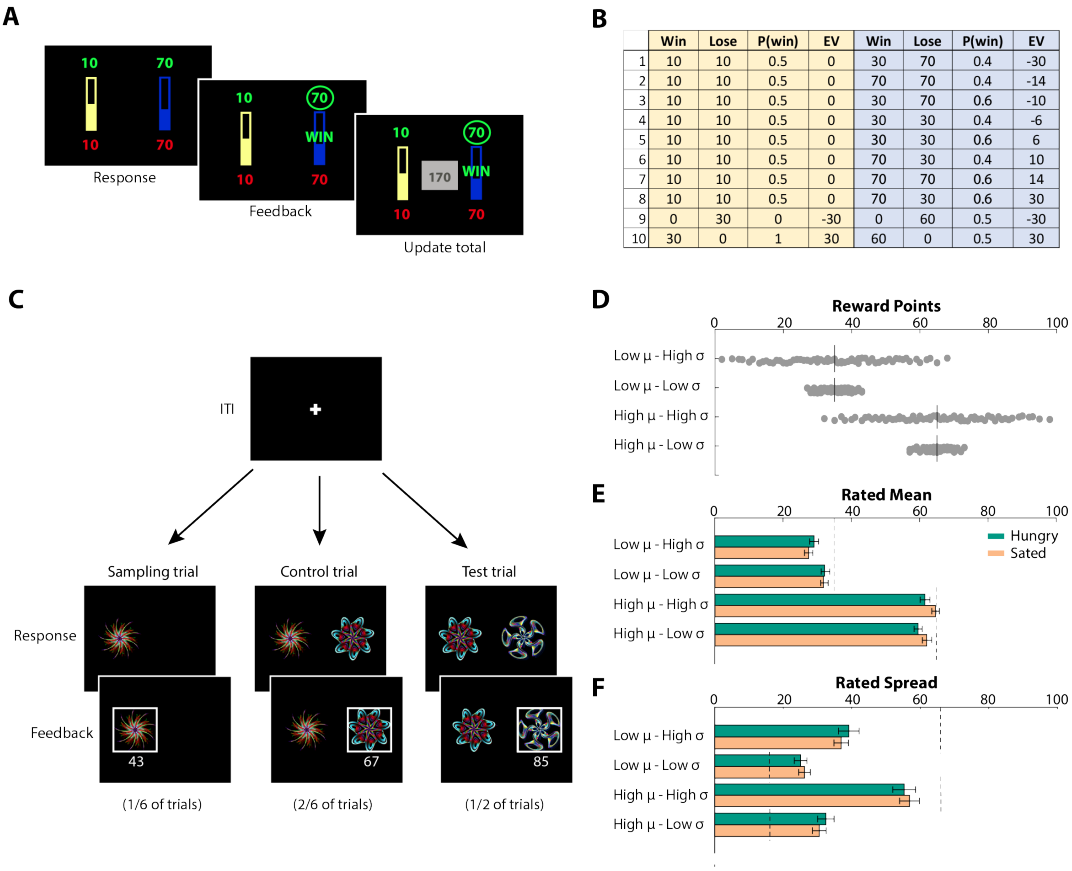
has mainly focussed on explicitly described lottery gambles without exploring the potential effects of food deprivation on positive and negative decision contexts. To gain a better understanding of how individual risk attitudes are modulated by how risks are presented, the decision context, and the level of food deprivation, I employed two complementary risk-taking tasks. One task involved decisions between two options of which the probability of winning and losing and the magnitude of rewards were explicitly described. The other task involved decisions between options of which the average reward and uncertainty had to be learned through sampling. In both tasks, participants were presented with options that differed in the level of risk and had either the same or different expected values. The combinations of options were categorised into the following three decision contexts: a mixed context in which the expected value of the two options were different; a negative context in which options had equal expected values below the reference point; and a positive context in which options had equal expected values above the reference point.

Following previous studies, I expected that food deprivation also affects risk-taking of monetary outcomes. In line with the description-experience gap, I expected that risk attitudes are weighted differently depending on whether the decision context is positive or negative and that these patterns are reversed for decisions from experience compared to decisions from description. Furthermore, I hypothesise that decisions from experience rely on the value signals encoded in the Go and Nogo pathways of the basal ganglia, which makes them vulnerable to dopamine modulation. As predicted by the theory, food deprivation may therefore increase overall risk-seeking for learned risks. The explicit evaluation of risks may not be affected by food deprivation, because this relies on cortical processes that may not be under the direct control of dopamine. Lastly, if prospects are under- or overweighted depending on the framing of prospects, food deprivation may also modulate risk-taking differently depending on the context.

6.2 Methods

Experimental design

Decisions by description Risk-taking behaviour in decisions from description was probed using the probabilistic task described by Rogers and colleagues (Rogers et al., 2003). On each trial, participants were required to choose between two simultaneously presented gambles (Fig. 6.1A). Each gamble was represented visually by a histogram of which the height indicated the relative probability of winning a given number of points. The magnitude of possible points to win was indicated in green above each histogram, with the magnitude of possible points to lose indicated below in red. I used 10 different types of gambles with different decision contexts (Fig. 6.1B). Eight of these gamble types offered a choice between two gambles that differed in their objective expected values (mixed decision context). The reference gamble always consisted of a 50:50 chance of winning or losing 10 points, giving an expected value of 0. The risky gamble varied in the probability of winning (0.6 or 0.4), magnitude of possible points to win (30 or 70 points), and magnitude of possible points to lose (30 or 70 points). The remaining two gamble types offered a choice between a certain win or loss and a 50:50 chance gamble with the same expected value. These gamble types were identical in terms of prospect, but differed in valence, allowing for the examination of positive and negative context effects on differences in risk attitudes. Visual feedback (win/lose) was given after each choice was made, and the revised running total points was presented before the next trial. Note that this was not a sampling task where the feedback was used to perform optimally. Every gamble was considered independently of outcomes of previous gambles. Participants completed four blocks of 20 trials, and the order in which gambles were presented was kept constant for both conditions. Participants were instructed that the highest total score obtained in a block would be converted into pence and paid at the end of the task as a performance bonus. Deliberation times were also recorded.



	Win	Lose	P(win)	EV	Win	Lose	P(win)	EV
1	10	10	0.5	0	30	70	0.4	-30
2	10	10	0.5	0	70	70	0.4	-14
3	10	10	0.5	0	30	70	0.6	-10
4	10	10	0.5	0	30	30	0.4	-6
5	10	10	0.5	0	30	30	0.6	6
6	10	10	0.5	0	70	30	0.4	10
7	10	10	0.5	0	70	70	0.6	14
8	10	10	0.5	0	70	30	0.6	30
9	0	30	0	-30	0	60	0.5	-30
10	30	0	1	30	60	0	0.5	30

Figure 6.1: **Overview of study design.** A) Example trial of the description-based risk-taking task. Each trial consisted of a choice between a reference gamble (yellow) and risky gamble (blue). Points to win and lose were presented in green and red, respectively. The probability of winning corresponds to the size of the filled bar. Feedback was given after a choice and the running total was updated. B) Gamble types 1–8 show different combinations of points to win, points to lose and the probabilities of winning, with expected values ranging from -30 to 30. Trial type 9 and 10 have equal expected values, but different risks. C) Task structure for experienced-based risk-taking. The task consisted of three different trial types: Sampling trials (1/6 of the trials), control trials (2/3 of the trials) and test trials (1/2 of the trials). After a response, a reward sampled from the associated reward distribution was presented. D) Each reward distributions associated with a stimulus was approximately normal. The mean of the distribution was either 35 (low μ) or 65 (high μ) with a standard deviation of either 5 (low σ) or 20 (high σ). Each grey dot indicated a sample drawn from the reward distribution. Subjectively rated values for mean (E) and spread (F). Dashed lines indicate the objective mean or spread of a reward distribution. Error bars represent SEM.

Decisions by experience Risk-taking behaviour in decisions from experience was probed using a partial-feedback design with three trial types: test trials, control trials and sampling trials (Fig. 6.1C) (Ludvig et al., 2014; Niv

et al., 2012). Control and test trials were alternative choice trials in which participants aimed to choose the option that maximised the total payout. In test trials, stimuli were matched for mean (i.e. they have the same objective expected value) but differed in variance. In control trials, stimuli differed in their objective expected values, but not necessarily in their variances. Control trials allowed for testing whether participants paid attention to their decisions and provided an objective measure of whether people learned which stimuli had a higher expected value. Sampling trials were forced choice trials in which only one stimulus was presented that had to be chosen to continue to the next trial. These trials ensured that all options were sampled from and that participants occasionally experienced reward contingencies that they did not prefer (Ludvig et al., 2014; Niv et al., 2012).

Each trial had the same structure. After a short ITI of 500-700 milliseconds, the stimuli were presented on the screen. Responses were made by pressing on the left or right arrow key of the keyboard to choose the left or right option, respectively. Choices were immediately followed by feedback for 1.5 seconds, showing the number of points won (Fig. 6.1D). Each stimulus was associated with a Gaussian reward distribution that followed a two-by-two design: high or low mean value (65 or 35 points) and high or low variance value (20 or 5) (Fig. 6.1D). This design allowed for the investigation of context effects on risk attitudes. The total accumulated points was continuously displayed at the top of the screen. Participants completed four blocks of 72 trials. Reward distributions were generated at the start of each block to ensure each block had the intended reward distribution and stimulus sets were reset after 2 blocks (or 144 trials). The trial order was pseudo-randomised to ensure that a stimulus presented in a sampling trial did not precede a test trial with the same stimulus to avoid priming of choices. Participants were instructed to maximise their total number of points, which would be converted into a monetary performance bonus at the end of the task. After each block, participants were asked to indicate the reward distribution of each of the stimuli by placing two cursors,

one for the minimum and one for the maximum reward in the distribution, on a Visual Analogue Scale ranging from 0 to 100 points. The rated spread was computed as the difference between the rated minimum and maximum of the reward distribution and the rated mean was taken as the average of the two values.

Behavioural analyses

In these tasks, risk was defined as the uncertainty in the possible outcomes of a decision, expressed as the variance of the associated reward distribution (Rothschild and Stiglitz, 1970). Risk attitudes captured how often an individual preferred the risky option in a given context. For experience-based decisions, risk preference was averaged over the second half of the trials (72 trials) of each stimulus set and compared using t-tests or ANOVAs as appropriate. Using only the second half of the trials allowed participants sufficient opportunity to learn the outcomes associated with each option, while providing a long enough sample to get a reliable measure of their risk preference (Ludvig et al., 2014; Niv et al., 2012). All trial types were equally distributed over the blocks. The average expected reward of all stimuli was 50, which served as the reference point in this task. Mixed decision contexts included control trials, which have an average expected value that is equal to the reference point. Choices on these trials should be driven by a difference in mean, not variance, causing risk neutral behaviour. Negative contexts were low mean test trials that have an average expected value below the reference point. Positive contexts were high mean test trials that have an expected value above the reference point.

For description-based decisions, risk preference was assessed as the proportion of risky gambles chosen. Mixed decision context gambles were gamble types 1–8, the negative decision context was gamble 9, and the positive decision context was gamble 10. Trials in the description-based task were independent of each other and did not require learning of outcomes. Therefore, all trials

contributed equally and all trials were included for the risk preference and compared using t-tests or ANOVAs as appropriate.

For the comparison of context effects on task type, only matched mean trials concerning positive or negative decision contexts were included.

Computational modelling

In the fitting procedure for experience-based decision-making, control and test trials were used to estimate all parameters. Sampling trials were included for the estimation of learning rates for the mean and variance of a stimulus using Eq. (6.1) and Eq. (6.2), respectively. However, sampling trials were excluded from the estimation of the softmax choice parameter and the risk parameters, due to the absence of a choice. The presence of only one stimulus makes the probability of choosing this stimulus one, and this would interfere with the parameter estimation. Initial values for Q and S were set to 50 and 5, respectively.

6.3 Results

I first examined the risk attitudes in decisions from experience. In this task, optimal decision-making required participants to first learn the expected values of the available options by sampling, and then use the learned values to choose the option that maximises total number of points.

Participants learned different reward distributions

I verified whether participants were able to differentiate stimuli based on the mean and variance of the associated reward distributions. The effect of mean on choice behaviour was implicitly measured in control trials. All participants performed on average above 90% on the control trials, and no participant performed below 60%, suggesting that they understood the distinction between

high and low mean stimuli. The level of food deprivation did not affect the accuracy on control trials [$t_{31} = 0.62$, $p = 0.543$].

To verify whether participants understood the difference in mean and variance of reward distributions, participants were asked to indicate the reward distribution of each stimulus using a Visual Analogue Scale. I compared the objective reward points (Fig. 6.1D) with the subjectively rated mean and spread of the reward distributions (Fig. 6.1E–F). The mean and spread indicated by participants showed a similar pattern as the objective values, showing that participants were consciously able to reliably distinguish stimuli based on their mean and spread. These two measures provided convincing evidence that the risk attitudes observed in this task were driven by real preferences and were not caused by an inability to differentiate stimuli based on the variability in reward outcomes.

Food deprivation altered learned risks, but not described risks

For the learned risks, I analysed choice behaviour on the test trials to evaluate risk-taking behaviours in negative and positive decision contexts (Fig. 6.2A). Participants were significantly more likely to choose the risky option in a positive decision context, but not a negative context (main effect of context [$F_{1,31} = 10.28$, $p < 0.003$]). The under- and overweighting in the negative and positive decision context, respectively, is consistent with previously reported risk attitudes for learned risks (Ludvig et al., 2014; Madan et al., 2015), and was also reflected by the subjective rating of stimulus values at the end of the task (interaction effect of mean and variance on subjective rating [$F_{1,31} = 9.91$, $p < 0.004$]; Fig. 6.1E). Crucially, food deprivation modulated risk-attitudes for positive and negative contexts in opposite manner (interaction effect of food deprivation and context [$F_{1,31} = 8.38$, $p < 0.007$]). A *post hoc* paired t-test revealed that this interaction effect was mainly driven by hunger decreasing risk-taking behaviour in the positive context [$t_{31} = 2.73$, $p = 0.010$], and not in

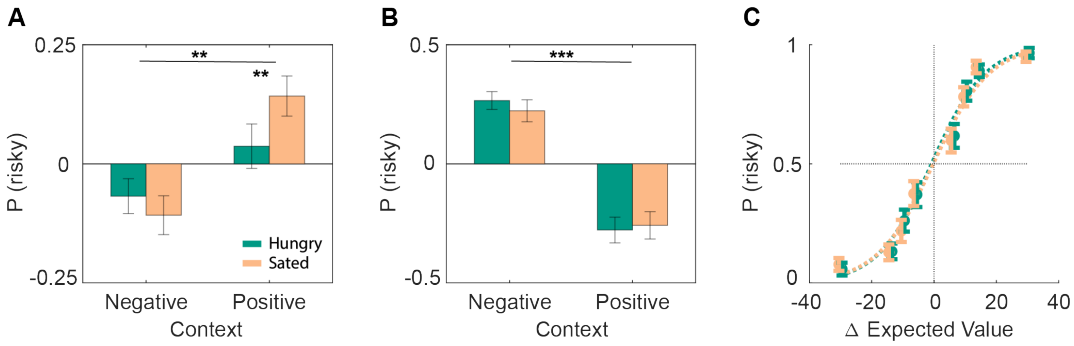


Figure 6.2: **Risk attitudes for learned and described risks.** A) In decisions from experience, participants were risk-averse for negatively framed choices and risk-seeking positively framed choices. Food deprivation attenuated risk attitudes for decision contexts in opposite direction. B) In decisions from description, participants were risk-seeking for losses and risk-averse for gains (gambles 9 and 10 Fig. 6.1E). Food deprivation did not affect these risk preferences. C) Proportion of risky gambles chosen for mixed context trials. Psychometric curves were estimated with a logistic model. The probability of choosing the risky gamble over the reference option increased with increasing positive difference in expected values and decreased with increasing negative difference in expected value (gambles 1–8 Fig. 6.1E). There was no difference in choice behaviour between hungry and sated participants. Error bars represent SEM. ** $p < 0.01$, *** $p < 0.001$.

the negative context [$t_{31} = 1.01$, $p = 0.319$]. Finally, food deprivation did not alter overall risk-taking behaviour (main effect of food deprivation [$F_{1,31} = 1.19$, $p = 0.283$]).

To provide a comparable measure to the context effects in experience-based risk-taking, I also analysed the risk preference for matched mean gambles in positive and negative decision contexts in description-based choices (Fig. 6.2B). Participants were significantly more likely to choose the risky gamble compared to a certain option when the expected value for both options was -30, making them more risk-seeking for negatively framed choices. In contrast, participants chose the risky gamble less often compared to the certain option when both options had an expected value of 30, making them risk-averse for positively framed choices (main effect of context [$F_{1,31} = 55.01$, $p < 0.0001$]). The over- and underweighting in the negative and positive decision context, respectively, has been previously described by prospect theory (Kahneman and Tversky, 1979). In contrast to learned risks, food deprivation did not alter context-

specific risk attitudes for described risks (interaction effect of food deprivation and context [$F_{1,31} = 1.53$, $p = 0.255$]) or overall risk-taking (main effect of food deprivation [$F_{1,31} = 0.13$, $p = 0.717$]; Fig. 6.2B).

For described risks in mixed decisions contexts, I found that participants were more likely to choose the risky gamble over the reference option when the difference in expected value was positive and less likely when the difference in expected value was negative (Fig. 6.2C). This suggests that participants were more (or less) willing to accept risks if this was reflected by a sufficient difference in expected value (Weber et al., 2004). Food deprivation did not attenuate the sensitivity or bias for choosing the risky gamble, suggesting that food deprivation did not shift the reference point for risky choices or impact how expected values were perceived (Fig. 6.2C). This observation is consistent with the performance of individuals on mixed context control trials in decision from experience, but is inconsistent with previous reports (Levy et al., 2013; Symmonds et al., 2010).

When comparing the risk attitudes across tasks, I found that risk attitudes were reversed by framing and by how risks were presented. These results suggest that risks may be perceived, or weighted, differently when they were described compared to when they were learned. The difference in risk attitudes for implicit and explicit risk-taking corresponded to a previously reported observation that is better known as the description-experience gap (Hertwig and Erev, 2009). Critically, food deprivation only affected the risk attitudes in decision-making from experience, but not from description.

Modelling of risk-sensitive choice behaviour

The previous analyses showed that food deprivation only altered decision-making when risks had to be learned. However, the behavioural analyses do not provide insight in what computational process was altered by food deprivation. Therefore, I employed a computational modelling strategy to account for the integration of a specific reward history triggered by sampling. This

strategy allowed me to attribute the effects of food deprivation to a specific computational process.

Most computational models that explain risk preferences in economic decision-making have been developed for decisions from description. Models such as the non-linear utility model (Kahneman and Tversky, 1979) and mean-variance model (Boorman and Sallet, 2009), function well for independent gambles, but may fall short when outcomes are learned. For experiential learning, trial-and-error reinforcement learning models have been proven invaluable (Sutton and Barto, 2017). I relied on two learning models to explain the behavioural data: a standard reinforcement learning model (Rescorla and Wagner, 1972) and a simplified version of the payoff-cost model (Mikhael and Bogacz, 2016; Möller et al., 2020), which I call the risk-sensitive model.

Standard reinforcement learning models provide trial-by-trial estimates of the expected mean value of each stimulus, without considering the variability in outcomes. They provide a good account of what a rational decision-maker would do based on the objective expected value of a reward, and choices are not evaluated differently for different decision contexts. In this model, the expected mean value of the chosen stimulus Q_c was updated using:

$$Q_{c,t+1} = Q_{c,t} + \alpha_Q(r_t - Q_{c,t}), \tag{6.1}$$

where r_t is the reward obtained by choosing a stimulus on trial t and α_Q is the learning rate for the mean reward. Decisions in this model were solely based on the expected mean value of the presented stimuli. The utility U of stimulus i on trial t was $U_{i,t} = Q_{i,t}$ and was entered into the softmax decision rule described in Eq. (2.14).

In contrast, the risk-sensitive model is an extended version of the standard Q-learning model that accounts for both the average outcome of an action (Q) and the variability, or spread (S), in outcomes. The rules for the expected mean and spread of reward distributions are analogous to the synaptic plasticity rules for the Go and Nogo pathways of the basal ganglia (Mikhael and Bogacz, 2016).

The mean of a distribution is encoded in the difference between the synaptic weights of these two pathways ($Q = G - N$), whereas the spread is encoded in the sum ($S = G + N$). These rules are consistent with the properties of dopaminergic receptors expressed in these pathways and have been shown to account for various empirical findings including the effects of dopamine on risk-taking. The risk-sensitive model captures innate risk propensities and assumes that positive and negative decision contexts influence how the spread in reward outcomes affects the subjective utility of an action. This model may provide a closer approximation of the heuristic approach humans could adopt when making context-dependent choices. Given the differential effects of positive and negative contexts on risk preference, this model is likely to explain the behavioural data better than a standard reinforcement learning model.

The mean expected value in the risk-sensitive model was also updated using Eq. (6.1). A measure of the spread in reward outcomes was learned in an analogous manner to Q-values:

$$S_{c,t+1} = S_{c,t} + \alpha_S(|r_t - Q_{c,t}| - S_{c,t}), \tag{6.2}$$

where α_S is the learning rate for the spread, and $r_t - Q_{c,t}$ is the reward prediction error that captures the deviation of the current outcome from the average outcome, which is compared with the current expected spread in reward outcomes $S_{c,t}$.

The risk-sensitive model accounts for how participants differentiate matched mean stimuli based on the spread and captures individual risk propensities. For this model, the utility that was entered into the softmax function, Eq. (2.14), depends on the expected mean reward, the spread in reward outcomes and the sensitivity to the decision context (i.e. the framing effect), in the following way:

$$U_{c,t} = \underbrace{Q_{c,t}}_{\text{Expected mean}} + \underbrace{\gamma_0 S_{c,t}}_{\text{Risk propensity}} + \underbrace{\gamma_1 \delta_{\text{context}} S_{c,t}}_{\text{Framing effect}}, \tag{6.3}$$

where the parameter γ_0 modulates the risk propensity of an individual. In the model by Mikhael and Bogacz (2016), γ_0 reflects the tonic level of dopamine. A

positive value of γ_0 increases risk-seeking as the size of the spread is positively contributing to the choice, meaning that the high-spread option would be preferred. This effect is reversed for when $\gamma_0 < 0$. Note that the first two terms in Eq. (6.3) are analogous to the mean-variance models developed for decisions from description (Boorman and Sallet, 2009; D'Acremont et al., 2009).

The third term, the framing effect, captures the biasing effect of positive or negative decision contexts on choice behaviour. Framing effects play an important modulatory role in risky decision-making (De Martino et al., 2006; Mishra et al., 2012; Palminteri et al., 2015; Tsetsos et al., 2012; Tversky and Kahneman, 1981) and were also observed in the current study. The context reflects how the expected value of the presented stimuli compares to the overall expected value of all stimuli in the task $\delta_{\text{context}} = Q_{\text{presented},t} - Q_{\text{all,t}}$, where $Q_{\text{presented},t}$ is the average of the Q-values of the stimuli on the current trial and $Q_{\text{all,t}}$ is the average of the Q-values of all four stimuli. The value of $Q_{\text{all,t}}$ represents the reference point in this task. The objective value of the reference point is 50 points, but it fluctuates around 50 as the Q-values of the stimuli change by trial-to-trial updates. Positive decision contexts concern high mean test trials that have an objective value above the reference point ($\delta^{+}_{\text{context}} = 65 - 50 = +15$ points), whereas negative decision contexts concern low mean test trials that have a context value below the reference point ($\delta^{-}_{\text{context}} = 35 - 50 = -15$ points). The parameter γ_1 is a gain parameter that determines the extent to which the decision context and spread in reward outcomes contribute to choice behaviour. Positive values of γ_1 increase risk-taking behaviour in positive contexts, and reduce risk-seeking in negative contexts. The opposite is true for negative values of γ_1. Critically, decision contexts change from trial-to-trial and may elicit positive and negative emotions (surprise or disappointment, receptively) when the presented context is better or worse than the average reward, or reference point, in the task (Möller et al., 2020).

In the risk-sensitive model, the effects of food deprivation can be attributed to how participants learn about the expected value (reflected by α_Q), the spread

of reward outcomes (reflected by α_S), the individual risk propensity (reflected by γ_0) and/or sensitivity to the framing effect (reflected by γ_1).

Computational modelling captured risk preferences

I used a hierarchical model fitting procedure as described in the Methods (section 4.3). The model parameters, denoted by Greek letters, were transformed into a Gaussian scale, denoted by their respective Latin letters, using Eq. (4.7) for the learning rates, Eq. (4.9) for the softmax parameter, and Eq. (4.8) for the risk-sensitive parameters. The statistical significance was tested with respect to the Gaussian scaled model parameters and corrected for multiple comparisons. A model comparison (see Methods section 4.3) revealed that over 70% (23 out of 32 participants) were better described by the risk-sensitive model ($\text{BIC}_{\text{RW}} = 16915$ and $\text{BIC}_{\text{risk-sensitive}} = 15996$), confirming that the addition of extra parameters was justified. The quality of the fitting procedure was verified with a parameter recovery analysis (see Methods section 4.3). All parameters were well recovered ($0.75 < R < 0.95$) and the model fitting procedure did not introduced spurious correlations between the other parameters ($|R| < 0.3$; Fig. 6.3A). Surrogate data generated with the best fitted parameters confirmed that the model reproduces the key effect of food deprivation on choice preferences (Fig. 6.3B).

In line with the behavioural analyses, I found an effect of food deprivation on behaviour fitted with the risk-sensitive model (Fig. 6.3C). On average, food deprived participants had lower learning rates for the spread [α_S, $p < 0.0001$] and a lower sensitivity to framing effects [γ_1, $p = 0.004$]. Food deprivation did not significantly alter learning rates for mean values [α_Q, $p = 0.033$], when corrected for multiple comparisons, or choice stochasticity [β, $p = 0.496$]. At the group level, individual risk propensities were not significantly altered by food deprivation [γ_0, $p = 0.595$]. The γ_0 parameter captures the individual differences in risk propensity as a function of the level of food deprivation. In line with previous findings (Levy et al., 2013), food deprivation does not have

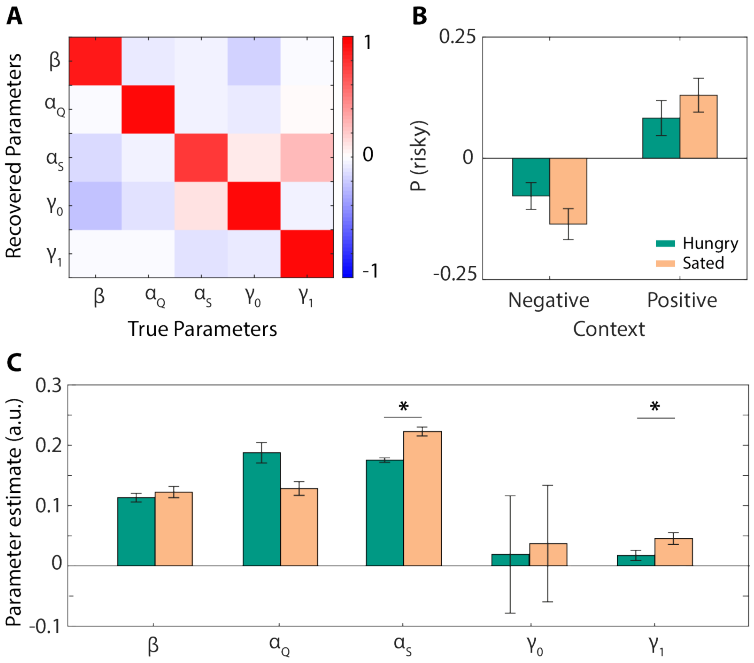


Figure 6.3: **Parameter estimates risk-sensitive model.** A) Correlation matrix of the free parameters used to generate the simulated data (True parameters) and the obtained parameters by applying the parameter estimation procedure on the simulated data (Recovered parameters). Bright red values indicate a strong correlation between the true and recovered parameter value and therefore a good parameter recovery. B) Simulated choice behaviour using estimated parameters for the food deprived and sated condition revealed a similar pattern as the behavioural data depicted in Fig. 6.2A. C) Food deprivation significantly decreased the learning rate for reward spread α_S and the sensitivity to contexts γ_1. The effect of food deprivation on the learning rate for mean α_Q, the softmax temperature β or individual risk preferences γ_0 was not significant. Error bars represent SEM. Statistical significance was tested with respect to the Gaussian scaled parameters. * $p < 0.05/5$ (corrected for multiple comparisons).

the same effect on risk attitude in all individuals. Together, these data show that hunger altered risk preferences by modulating the learning rate for spread and the sensitivity to framing effects.

Individual risk propensities and the sensitivity to decision-contexts were captured by individual parameters in the model. The effect of food deprivation on overall risk-taking (i.e. the probability of choosing the risky option for both positive and negative decision contexts) was captured by the change in the γ_0 parameter derived from computational modelling ($[R = 0.81, p < 0.0001]$; Fig. 6.4A). The effect of food deprivation on the context effects was captured

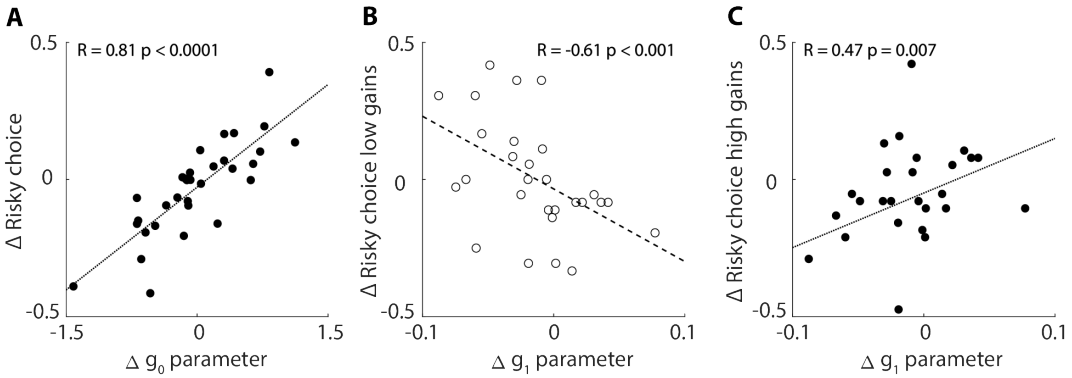


Figure 6.4: **Changes in risk attitudes were captured by computational modelling.** A) The shift in sensitivity to variance (Gaussian scaled g_0 parameter) correlated positively with the change in overall risk-seeking behaviour individuals (i.e. risk-seeking behaviour for high and low gains). B) Change in risk attitude for low means was negatively correlated with the shift in sensitivity to the context (Gaussian scaled g_1 parameter) C) Change in risk attitude for high means was positively correlated with the shift in sensitivity to the context (Gaussian scaled g_1 parameter). Positive values indicate a higher value in the food deprived condition.

by the change in γ_1. Changes in risk-taking observed for negative (Fig. 6.4B) or positive decision contexts (Fig. 6.4C) correlated strongly with the change in γ_1, which is in line with the asymmetrical modulation by food deprivation seen in Fig. 6.2A.

Using the best-fitting parameters of the risk-sensitive model, I estimated the latent mean (Q) and spread (S) of the different stimuli for both conditions (Fig. 6.5). Estimates were averaged across stimulus sets and participants. The grey shading in Fig. 6.5 indicates the trials that were used for the behavioural analyses depicted in Fig. 6.2A. The estimates derived from computational modelling showed that learning had stabilised at this point, providing additional evidence that the measured risk preference was not caused by differences in learned expected value or spread between conditions or stimuli. These data also confirm that participants were able to differentiate stimuli based on the mean and spread of the associated reward distribution, which is in line with subjective rating (Fig. 6.1E–F).

Learned mean (Q) values corresponded well to their respective objective expected values, indicated by the asterisks in Fig. 6.5. The learned spread (S)

of a distribution differed from its objective value, but the learned value for high and low variance stimuli was clearly separated. Comparison of food deprived and sated conditions showed that food deprivation did not alter the asymptotic value for the mean and spread of the four reward distributions.

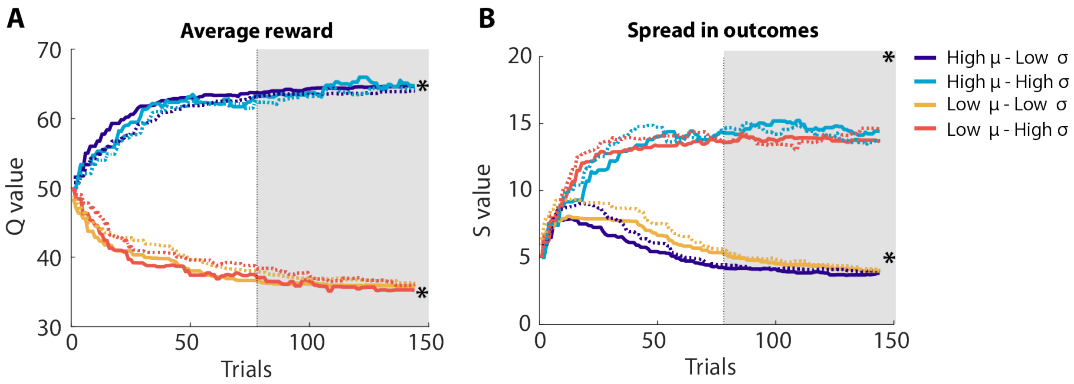


Figure 6.5: **Learned mean and spread of reward distributions.** Learned mean (Q) and spread (S) were derived from the risk-sensitive model using the best fitting parameters. Food deprivation did not affect the asymptotic values for the average reward (A) or spread in reward outcomes (B). Values were averaged across stimulus sets and across participants. Solid line represent food deprived individuals and dashed lines represent sated individuals (Fig. 6.2A). Grey shaded area indicated the trials that were used in the risk preference analysis. * denotes the true value of the mean or spread.

Reaction times did not explain changes in risk attitudes

Finally, I examined whether the effect of food deprivation on risk-seeking behaviour, or lack thereof, was the result of changes in reaction times. Food deprivation did not affect overall reaction times in decisions made from experience (main effect of food deprivation [$F_{1,31} < 1$]; Fig. 6.6A). Participants changed their reaction times depending on the decision context (main effect of decision context [$F_{1,31} = 37.67$, $p < 0.0001$]). Further analyses using a two-way repeated measures ANOVA revealed that participants took significantly more time to respond for negative decision contexts compared to positive contexts [$F_{1,31} = 51.39$, $p < 0.0001$] and mixed contexts [$F_{1,31} = 39.19$, $p < 0.0001$], without showing a significant difference in reaction times between positive contexts and mixed contexts [$F_{1,31} = 0.55$, $p = 0.464$]. These data suggest that

negative contexts were perceived as more difficult than positive or mixed contexts, a finding that has been reported before (Madan et al., 2015). Food deprivation did not affect context-specific reaction times (interaction effect of food deprivation and context [$F_{1,31} < 0$]).

For decisions from description, food deprivation increased reaction times for all gambles types, regardless of the decision context ([$t_{31} = 4.57$, $p < 0.0001$]; Fig. 6.6B). Despite the absence of an effect of food deprivation on choice behaviour, participants took more time to decide whether to gamble or not.

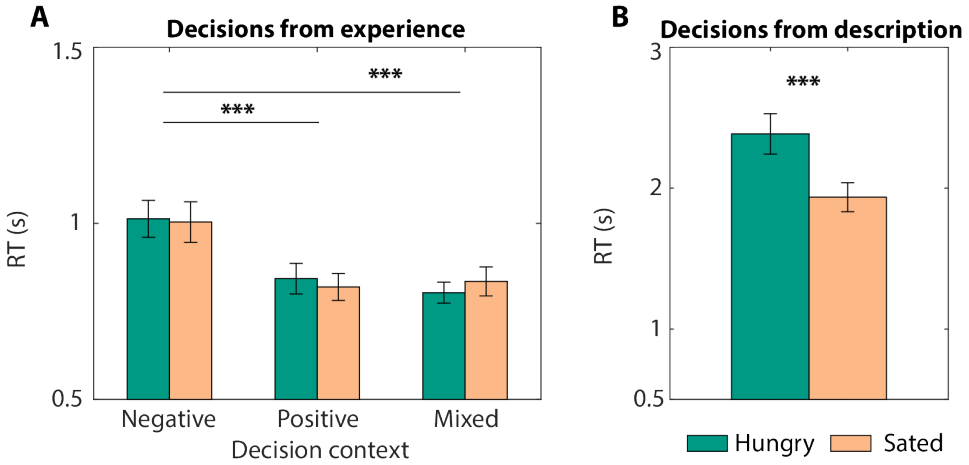


Figure 6.6: **Food deprivation increased reaction times for described risks only.** A) In decisions from experience, median reaction times for the negative decision context were higher than for the positive and mixed decision context. There was no effect of food deprivation on reaction times. B) In decisions from description, reaction times were increased for the food deprived condition. Error bars represent SEM. *** $p < 0.001$.

6.4 Discussion

The integration of information about the current metabolic state and the ability to adapt to variable reward outcomes is critical for survival (Stephens, 1981). In this study, I used two complementary tasks to test whether food deprivation selectively affected risk attitudes depending on whether risks were learned or explicitly described and whether the outcomes were framed in a positive or

negative context. I found that food deprivation modulated risk attitudes for decisions of which outcome statistics had to be learned, but not for decisions of which outcome statistics were explicitly described. In decisions from experience, food deprivation enhanced risk-seeking for negatively framed choices and promoted risk aversion for positively framed choices. These results suggest that the current metabolic state drives adaptive behaviour for trial-and-error learning in a context-specific manner, but may not alter the integration of factual information.

Experience versus description

Research on decision-making under uncertainty has revealed two distinct modes of decision-making that are differentiated based on how risks are presented. Decisions could be made based on the explicit knowledge of outcomes statistics, or based on the learned outcomes statistics through sampling. Following prospect theory (Kahneman and Tversky, 1979), participants were risk-seeking in decisions concerning losses, and risk-averse in decision concerning gains. In line with previous studies involving experiential learning of uncertain outcomes (Hertwig et al., 2004; Ludvig et al., 2014; Ludvig and Spetch, 2011; Madan et al., 2014; Tsetsos et al., 2012), I found that participants were more risk-averse for negative decision contexts and risk-seeking for positive decision contexts. These opposing patterns of risk-seeking have been reported before, but not in the same individual. These data suggest that the way risks are presented, rather than individual risk propensities, play an important role in this observation.

The underlying reason for different risk patterns for described and learned risk are thought to come from underweighting or overweighting of rare events, respectively (Hertwig, 2012; Hertwig et al., 2004; Kahneman and Tversky, 1979), or memory biases (Madan et al., 2014). Particularly, when risks need to be learned, sampling of reward outcomes may alter preferences (Camilleri and

Newell, 2011; Hertwig and Erev, 2009). To ensure that risk attitudes in decisions from experience were not driven by under-sampling of options, I added forced sampling trials to the partial feedback design. Participants occasionally experienced outcomes of stimuli they did not prefer. Risk attitudes were based on choice behaviour once learning had stabilised (Ludvig et al., 2014; Ludvig and Spetch, 2011; Niv et al., 2012). Ratings of reward distributions and computational modelling suggest that risk preference are most likely the result of differences in sensitivity to decision contexts, rather than differences in learning.

Despite the opposing risk patterns observed in decisions from experience and from description, in both tasks, decisions were evaluated with respect to a reference point that is specific to the individual and the situation (Tversky and Kahneman, 1981). Regardless of whether the reference point was zero, positive or negative, the evaluation of decisions with respect to a reference point seems to produce similar patterns of risk-taking (Ludvig et al., 2014). This suggests that decision contexts across tasks could be compared even when the prospects of positively and negatively framed decisions may differ as seen in this study.

Effects of food deprivation

Food deprivation only altered risk attitudes for decisions from experience, but not for decisions from description. This result fits with the idea that implicit risk-taking relies on the basal ganglia and is modulated by the current metabolic state. The absence of an effect on explicit risk-taking may be surprising, because previous studies showed that hunger increased risk-seeking for food, water and monetary rewards when gambles were explicitly described (Levy et al., 2013; Shabat-Simon et al., 2018; Symmonds et al., 2010).

Explicit risk-taking The absence of an effect in the current task could be due to differences in task design. The current study included only 10 sets of unique gambles that were played 8 times each, whereas previous studies

included around 150 – 200 unique gambles that differed in objective expected value and probability of winning, and gambles were only played once. Many of these gambles concerned a decision between a fixed certain reference option and a risky option that differed in expected value and uncertainty (Levy et al., 2013; Shabat-Simon et al., 2018). Other gambles involved a decision between options that varied in risk and expected value (Symmonds et al., 2010). These studies did not differentiate between decision contexts, making a direct comparison between the current data and previous studies difficult.

Another important difference between the current study and previous studies is the presence of feedback. Most economic decision-making tasks omit feedback, because all prospects/gambles should be considered independently of each other and experience with a lottery does not alter the outcome statistics of a described lottery. One study showed that the risk pattern predicted by prospect theory can be flipped by providing feedback of explicitly described gambles in a task that did not involve any learning (Jessup and O'Doherty, 2010). Feedback in the current task did not seem to have this effect, suggesting that the presence of feedback did not change the evaluation of gambles. Despite clear instructions before the task, it is possible that the presentation of the running total after each gamble impacted how participants evaluated upcoming choices. The running total of points could be perceived as the current wealth of an individual, which could alter risk-taking depending on whether the decision context is perceived as positive or negative. After the task, some participants reported that they changed strategies after experiencing multiple consecutive losses, especially if the running total became negative. This type of behaviour may be caused by a discrepancy between the reference point and the current number of points resulting from a shift in reference points to which the individual has not adapted yet (Markowitz, 1952), and may have contributed to the absence of an effect.

Implicit risk-taking Limited research is available on the effect of food deprivation on risk-taking for learned risks in humans. One study investigated the effects of hunger on risk-taking in an Iowa gambling task. In this task, participants choose between four deck of cards that follow a two-by-two design (analogous to the design in the current study): two decks have a positive expected value and two decks have a negative expected value. The decks with equal expected value differed in risk. They found that hunger enhanced the selection of high outcome options. However, this study did not investigate differences in preference for options that differed in risk, making a comparison with the current study difficult.

The effect of food deprivation in my study shows some resemblance to the risk attitudes described by foraging theory. Organisms tend to become less risk-averse (and even risk-seeking) under extreme hunger conditions, because they need to take more risks in order to find food for survival. The effects of food deprivation are also influenced by the decision context, making humans more or less conservative in their decision-making. Although the participants in this study were not starving, the tendency of hungry individuals to alter risk-taking is evidence of how evolutionary pressure from the past is still influencing our behaviour.

Considerations A couple of other considerations need to be made for the comparison with other studies. There is great variability in risk preferences among individuals, an observation that I also found in the current study. The effect of food deprivation may have different effects depending on the innate risk propensity of an individual. Previous studies showed that hunger had a converging effect on a population individuals who were highly risk-averse when satiated became less so when hungry, while risk-seeking individuals became more risk-averse (Levy et al., 2013). Another consideration is that the level of hunger may not be the same in all studies. Based on physiological measurements, risk preferences seem to correlate and depend on the severity

of the deprivation (some studies report 4 hours deprivation, others 12 hours) (Shabat-Simon et al., 2018; Symmonds et al., 2010). The current data could be very valuable, because I compared risk attitudes for described and learned risks in the same individual following the same level of deprivation.

Neural processing of risky decision-making

Decision-making is grounded in the anticipation and evaluation of prospective gains and losses. The identified neural systems underpinning these processes, particularly for uncertain outcomes, may be difficult to interpret due to the wide variety of task designs used for economic decision-making. Processing of information about uncertain rewards has been linked to many brain regions, including the frontal cortex, basal ganglia, amygdala, parietal cortex, cingulate cortex, and insular cortex (Christopoulos et al., 2009; Fiorillo et al., 2003; Huettel et al., 2006; Knoch et al., 2006; Kuhnen and Knutson, 2005; Levy et al., 2010; McCoy and Platt, 2005; Preuschoff et al., 2006; Tobler et al., 2009; Xue et al., 2009). In this section, I will address the systems involved in implicit and explicit risk-taking separately. I also discuss the possibility that risk and learning/evaluation of reward values may be represented separately in the brain, in an analogous manner to how economic theories separate risk and utility (Markowitz, 1952; Weber et al., 2004). Finally, I address the additional effects of framing (De Martino et al., 2006).

Implicit risk-taking Dopamine modulates risk-taking in PD patients (Gallagher et al., 2007) and healthy humans (Rigoli et al., 2016). Dopamine-rich brain areas such as the striatum and midbrain structures are sensitive to the expected reward (or utility) in learned and described risks (Abler et al., 2006; Hsu et al., 2009; Knutson et al., 2001; Niv et al., 2012; Tobler et al., 2007).

I hypothesised that implicit risk-taking may depend on the values encoded in the Go and Nogo neurons of the basal ganglia. I assumed that food deprivation increases the dopamine activation signal, which promotes the Go pathway

and inhibits the Nogo pathway. The differential effects of dopamine on the response to rewards and punishments (Cools et al., 2009; Frank et al., 2004) could encourage risk-taking by outweighing previous poor outcomes with recent successful outcomes. Consequently, gambling tends to be seen as a disorder of reward and punishment processing. Empirical data indeed suggest that changes in activity in the Nogo pathway contribute to risk preferences, although this effect may be specific to the target brain area and the type of reward outcome. Activation of NAc Nogo neurons biases choice away from riskier options (Zalocusky et al., 2016), whereas activation of Nogo neurons in the dorsal striatum promotes continued choice of options associated with potential loss (Tai et al., 2012). This suggests that the activity of Nogo neurons may play an important role in shifting preferences, particularly following recent losses, which could play a role in the differential effects on risk attitudes observed in positive and negative decision contexts.

In the model by Mikhael and Bogacz (2016), the parameter γ_0 modulates the risk propensity of an individual and reflects the dopamine activation signal. As expected, I showed that an increase in the γ_0 parameter correlates with an increase in overall risk-seeking behaviour. It will be interesting to investigate whether the γ_0 parameter also correlates with changes in hormone levels or neural activity as assumed in the model.

Dopamine also plays a crucial role in reward-based learning and encodes the reward prediction error (Schultz et al., 1997), which is modulated by the level of food deprivation (Aitken et al., 2016; Cone et al., 2016; Papageorgiou et al., 2016). Food deprivation may therefore affect the early stages of when probabilistic contingencies are being acquired. Computational modelling showed that food deprivation decreased the learning rate for spread, without altering the final learned values, and had no effect on learning the average reward. It is possible that food deprivation alters the sensitivity to positive and negative reward prediction errors, but this does not seem to alter overall risk preferences.

In addition to the neural responses found in subcortical areas, including the basal ganglia, implicit risk-taking has also been shown to evoke neural responses in cortical areas, such as the insular cortex (D'Acremont et al., 2009; Niv et al., 2012). This suggests that implicit risk-taking may engage the model-free and model-based system and that the trade-off between these two systems may depend on how risks are presented.

Explicit risk-taking Although some evidence suggest a role for the striatum in the evaluation of explicit risks (Hsu et al., 2009), most studies implicate cortical areas such as insula, posterior cingulate cortex, the orbitofrontal cortex and medial prefrontal cortex in tracking and representing risk in risk-sensitive decisions (Clark et al., 2008; Elliott et al., 1999; Hsu et al., 2005; Huettel et al., 2005, 2006; Knutson and Bossaerts, 2007; Kuhnen and Knutson, 2005; McCoy and Platt, 2005; O'Neill and Schultz, 2010; Platt and Huettel, 2008; Preuschoff et al., 2008; St. Onge et al., 2011; Tobler et al., 2007). These studies and the mechanisms they propose for the incorporation of risk into choice may depend on the model-based system. The theory in chapter 3 does not make any prediction about the effects of food deprivation on cortical areas. Previous studies showed that explicit risk evaluation may also be controlled by food deprivation, but these observations have not been linked to specific neural processes yet. If food deprivation indeed increases the dopaminergic activation signal it is possible that it exerts motivational control over goal-directed actions via its effect on dopamine in the cortex (Moscarello et al., 2007, 2009). Future studies will be necessary to investigate whether this is indeed the case.

Framing effects The behavioural data showed that the decision context was important for choice behaviour. For example, participants preferred the risky option in positive decision contexts, but preferred the safer option when it was presented with a low mean stimulus in a control trial.

Risk-taking for positively and negatively framed decisions may be driving by anticipatory emotions (De Martino et al., 2006). Positive aroused feelings associated with anticipation of gain (e.g. excitement/surprise) engage the VS (Knutson et al., 2001; Tobler et al., 2007) and may promote risk-taking. Negative aroused feelings associated with anticipation of loss (e.g. anxiety/disappointment) engage the amygdala (Canessa et al., 2013; Yacubian et al., 2006) and may promote risk aversion.

Although most of these studies involve described risk, recently a similar finding has been reported for learned risks (Möller et al., 2020). Positive and negative decision contexts were linked to phasic changes in pupil dilation. These data suggest that a change in pupil dilation for a positive decision context may encode surprise, reflecting that the potential prospect of this context may be better than expected. Changes in pupil dilation for a negative decision context may reflect disappointment in the potential prospect, because it is lower than the average reward in the task. These phasic changes also correlated to the γ_1 parameter, which captures the framing effect for different decision contexts. Where in the brain this framing effect is encoded for decisions from experience will need to be investigated in future studies. It may be possible that positive and negative decision contexts carry some sort of prediction error themselves, and interact with the dopamine (Preuschoff et al., 2006) or noradrenaline system (Preuschoff et al., 2011). Metabolic signals could amplify or reduce these effects via their control of dopamine or any downstream targets.

Conclusion

In conclusion, I found that food deprivation increased risk-taking for negatively framed choices, and decreased risk-taking for positively framed choices in decisions where outcome statistics had to be learned. This observation resembles predictions made by foraging theories, which show that survival rates increase when individuals consider the variability of resources in the environment according to the current level of energy reserves. Food deprivation did

not alter risk preferences for explicitly described risks, suggesting that cognitive evaluation of risk may be unaffected. This is the first study that uses a within-subject design to test the effects of food deprivation on risk attitudes for decisions involving learned and described risks in positive and negative decision contexts. It provides new insights into the modulatory role of hunger in adaptive behaviour. Further studies will be necessary to unify the data on the neural processing of decision-making under uncertainty. These studies will need to standardise task designs to allow for comparison between learned and described risks. Ideally, these tasks will also allow for the separation of risk evaluation and choice selection to identify neural systems that underpin risky decision-making and examine the impact of food deprivation on these processes.

Planning is bringing the future into the present so that you can do something about it now.

— Alan Lakein

7

The influence of metabolic state on learning strategies

Contents

7.1 Introduction 161

7.2 Methods 163

Paradigm 163

Stay-Switch Behaviour 164

Computational modelling 165

7.3 Results 167

Food deprivation only modulated model-free control 167

Food deprivation enhanced performance 170

Other reinforcement-driven behaviours were not affected . . 170

Computational modelling confirmed model-free effects . . . 172

7.4 Discussion 176

Learning strategies 176

Neural substrates of model-free and model-based control . . 178

Limitations 180

Conclusion 181

7.1 Introduction

The ability to plan ahead and satisfy any physiological need that arises is crucial for survival and allows individuals to avoid deviating too far from a balanced physiological state (Friston, 2010; Sterling, 2012). In chapter 3, I assumed that individuals reliably infer their current energy reserves and discussed how individuals could plan their actions in such a way to maintain or restore a physiological balance by creating an internal model of the body in the world (Seth and Friston, 2016). Hunger seems to have two contradictory functions on decision-making: on one hand, food deprivation may enhance adaptive behaviour by planning ahead (Higgs et al., 2017), on the other hand food deprivation promotes reactive behaviour by increasing impulsivity and reward sensitivity to encourage individuals to act fast and ignore potential negative consequences to ensure survival (Bartholdy et al., 2016; Kirk and Logue, 1997; Skrynka and Vincent, 2019). Previous research has focussed on the effect of food deprivation on only one of these two decision-making strategies, and has never directly compared the two in the same task. In this chapter, I aim to answer the following question: Does food deprivation promote simple reflexive decisions or prospectively-planned actions when both strategies can be employed simultaneously?

A frequently used paradigm to test the trade-off between reflexive and prospective decision-making is the two-stage sequential decision-making task developed by Daw et al. (2011). In this task, people must choose between two rockets, each going predominantly to one of the second-stage planets (Fig 7.1A). Once on a planet, they choose between two aliens which probabilistically deliver a reward. If a reward is obtained, it pays off to return to that planet on the next trial by selecting the same first-stage rocket again. Crucially, the rockets occasionally go to the other planet (a rare transition). If a reward is earned in this situation, it pays off to choose the *other* rocket next time, because that rocket is more likely to bring you back to the same second-

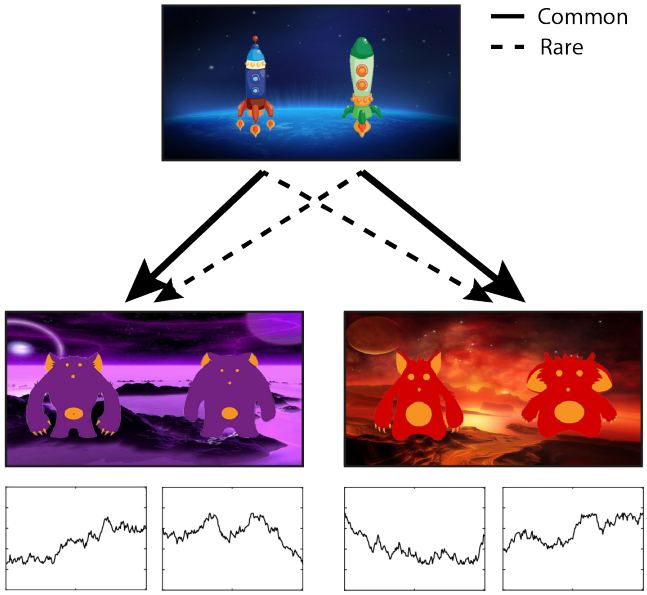


Figure 7.1: **Design sequential decision-making task.** Schematic of the task design developed by Daw et al. (2011). Each first-stage choice rocket flew predominantly to one of the second-stage planets (70% of the trials) and sometimes to the other second-stage planet (30% of the trials). Each second-stage alien probabilistically let to a reward. Reward probabilities changed slowly and independently throughout the task according to a Gaussian random walk ($\mu = 0$, $\sigma = 0.025$) with reflecting boundaries at 0.25 and 0.75. Task instructions and images were obtained from Kool et al. (2016).

stage planet. The simple, model-free approach is to repeat the rocket choice whenever it was rewarded. A more sophisticated, model-based strategy is to consider that rockets do not always go to the same planet and plan ahead to reach the most rewarding planet.

The arbitration of model-based and model-free behaviour is under the control of dopamine (Lee et al., 2014). Most studies have shown that increased levels of dopamine enhance model-based control (Deserno et al., 2015; Sharp et al., 2016; Wunderlich et al., 2012), but it is possible that other decision-making processes may be affected too (Kroemer et al., 2019). Striatal dopamine function is most clearly involved in reflexive, model-free strategies and prefrontal dopamine function is most clearly involved in flexible model-based strategies (Doll et al., 2016). Food deprivation increases dopamine levels within the basal ganglia (Aitken et al., 2016; Cone et al., 2016; Papageorgiou et al., 2016) and

dopamine utilisation in the mPFC (Carlson et al., 1987), suggesting that food deprivation could impact either of these strategies. Homeostatic hormones may be involved in cognitive control in the prefrontal cortex (Higgs et al., 2017), and cognitive control is impaired in obesity (Hassenstab et al., 2012) but enhanced in anorexia nervosa (Compan et al., 2015). Alternatively, food deprivation acts as a mild stressor (Deroche et al., 1995). Stress has been shown to impair model-based control (Otto et al., 2013) and enhances the reliance on model-free strategies, particularly for negative outcomes (Park et al., 2017).

Optimal behaviour on this task requires both simple reinforcement learning and prospective planning (Kool et al., 2016). If food deprivation increases impulsivity and reward sensitivity, behaviour could become maladaptive. Overall performance may be decreased, because relying too much on the model-free system could interfere with long-term goals. Alternatively, if food deprivation enhances model-based decision-making, overall performance may be increased, because actions are planned in advance to reach the most rewarding planet. In this study, I directly compare these two strategies and evaluate whether food deprivation enhances the reliance of one of the strategies, and promotes or impedes adaptive behaviour. This study provides insight into how changes in metabolic state can exert a generalised effect on decision-making strategies for monetary rewards.

7.2 Methods

Paradigm

In the two-step task, participants make a series of choices between two stimuli, which lead probabilistically to one of two second-stage states (Fig. 7.1). Each first-stage rocket leads more frequently (70%) to one of the second-stage states (a common transition), and in a minority of the choices (30%) to the other second-stage state (a rare transition). These second-stage states require a

choice between two aliens that offer different probabilities of obtaining a monetary reward. To encourage learning, the reward probabilities of these second-stage choices change slowly and independently throughout the task, according to a Gaussian random walk ($\mu = 0$, $\sigma = 0.025$) with reflecting boundaries at 0.25 and 0.75. The task consisted of 201 trials, divided in three blocks of 67 trials and separated by short breaks. If participants failed to enter a choice for a first or second-stage choice within 2 seconds, the trial was aborted and not included in further analyses. A new set of randomly drifting reward distributions was generated for each participant. Participants received 1 pence for every point they earned.

Before completing the full task, participants were extensively trained on different aspects of the task. Participants sampled from aliens with different reward probabilities, and were instructed on the transition structure (Kool et al., 2016). Finally, participants practised the full task for 25 trials. There was no response deadline for any of the sections of the training phase. Different rockets, planet colours and aliens were used for the training and experimental phase, and these were counterbalanced across conditions and participants.

Stay-Switch Behaviour

The one-trial-back stay-switch analysis is the most widely used method for characterising behaviour on the two-step task. This method quantifies the tendency of a participant to repeat the choice made on the last trial or switch to the other choice, as a function of the outcome and transition on the previous trial. I considered four possible outcomes: Common-Reward (CR), Rare-Reward (RR), Common-Unrewarded (CU), and Rare-Unrewarded (RU). Model-based and model-free indices were computed from the stay probabilities following each outcome according to:

$$\text{MF} = (P(\text{stay}|CR) + P(\text{stay}|RR)) - (P(\text{stay}|CU) + P(\text{stay}|RU)), \quad (7.1)$$

$$\text{MB} = (P(\text{stay}|CR) + P(\text{stay}|RU)) - (P(\text{stay}|CU) + P(\text{stay}|RR)). \quad (7.2)$$

Computational modelling

I used the original hybrid model by Daw et al. (2011), which assumes that there is an explicit trade-off between the model-based and model-free system. Model-based and model-free algorithms learn the value of the stimuli that appear in the task in three different pairs. There is one first-stage pair ($s1 \in \{1, 2\}$) and two second-stage pairs consisting of two stimuli each ($s2 \in \{3, 4, 5, 6\}$). The indices $s1$ and $s2$ refer to stage 1 and stage 2, respectively. The index t refers to the trial number.

At the first-stage, model-free 'cached' values were updated using the SARSA (λ) temporal difference algorithm. This algorithm learns to maximise the total outcome by strengthening or weakening associations between the first-stage state and the first-stage actions, depending on whether a reward followed that action or not:

$$Q_{s1}^{\mathrm{MF}}(t+1) = Q_{s1}^{\mathrm{MF}}(t) + \alpha_1(Q_{s2}(t) - Q_{s1}^{\mathrm{MF}}(t)) + \alpha_1\lambda(r - Q_{s2}^{\mathrm{MF}}(t)), \qquad (7.3)$$

where α_1 is the learning rate for the first-stage. The parameter λ is a gain parameter for the eligibility traces, which connects the two stages and allows the second-stage reward prediction error to influence first-stage choices. If $\lambda = 1$, the reward on stage 2 drives the choice on stage 1, and if $\lambda = 0$, outcomes on stage 2 do not influence stage 1 choices, which is representative of a pure model-based approach. The parameter λ also accounts for the main effect of reward as observed in the analysis of first-stage stay-switch behaviour, but not for an interaction of reward and state.

Model-based values were calculated for each first-stage stimulus and every trial in a forward looking manner by multiplying the state value of the best second-stage option with the state transition probabilities:

$$Q_1^{\mathrm{MB}} = 0.7 \times \max(Q_3, Q_4) + 0.3 \times \max(Q_5, Q_6), \qquad (7.4)$$

$$Q_2^{\mathrm{MB}} = 0.3 \times \max(Q_3, Q_4) + 0.7 \times \max(Q_5, Q_6). \qquad (7.5)$$

Model-based learning was simplified and the transition probabilities were not updated by explicitly modelling state prediction errors. Simulations by the authors of the original task showed that learning of state transitions quickly converge to stable values (see supplementary materials Daw et al. (2011)).

The hybrid model computes the actual value that is used in determining the stage 1 choice as a weighted combination of the model-based (Q^{MB}) and model-free (Q^{MF}) values. The first-stage Q-values were computed in the following way:

$$Q_{s1}^{\mathrm{hybrid}} = \omega Q_{s1}^{\mathrm{MB}} + (1-\omega) Q_{s1}^{\mathrm{MF}}, \tag{7.6}$$

where ω is a weighting parameter for the model-based strategy and $(1-\omega)$ is the weighting factor for the model-free strategy.

Q-values for the four second-stage stimuli were updated according to the reward prediction errors (Rummery and Niranjan, 1994):

$$Q_{s2}(t+1) = Q_{s2}(t) + \alpha_2 (r - Q_{s2}(t)), \tag{7.7}$$

where α_2 is the learning rate for the second-stage.
A first-stage choice depends on the relative difference in stimulus values between Q_1 and Q_2 and the choice C on the previous trial, which takes on the value 1 when the current choice equals the previous choice. The parameter π captures first-stage choice perseverance. Using the softmax choice function, the probability of choosing a first-stage stimulus was computed according to:

$$P_1 = 1/(1 + \exp(-\beta_1 (Q_{s1,1}^{\mathrm{hybrid}} - Q_{s1,2}^{\mathrm{hybrid}}) - \pi C)), \tag{7.8}$$

and for the second-stage:

$$P_3 = 1/(1 + \exp(-\beta_2 (Q_{s2,3} - Q_{s2,4})), \tag{7.9}$$

where β_1 and β_2 control the randomness of first and second-stage choices, respectively.

7.3 Results

I first examined the influence of food deprivation on choice behaviour. During each session, a participant undertook 201 trials, of which on average 2 trials were not completed due to failure to enter a response within the 2 seconds time limit. Food deprivation did not affect the number of missed trials [$t_{31} = 1.19$, $p = 0.245$] or median response times for first and second-stage choices (first-stage $\mathrm{RT}_{\mathrm{hungry}} = 616$ ms, $\mathrm{RT}_{\mathrm{sated}} = 640$ ms; paired t-test, [$t_{31} = -1.29$, $p = 0.207$], second-stage $\mathrm{RT}_{\mathrm{hungry}} = 788$ ms, $\mathrm{RT}_{\mathrm{sated}} = 821$ ms; paired t-test, [$t_{31} = -1.25$, $p = 0.220$]), suggesting that food deprivation did not alter overall attention in this task.

Food deprivation only modulated model-free control

Model-free and model-based strategies predict the following patterns of stay behaviour. A model-free reinforcement learning strategy predicts a main effect of reward (Fig. 7.2A), whereas a model-based learning strategy predicts a crossover interaction between reward outcome on the second-stage and the type of transition (Fig. 7.2B). The model-free and model-based system predict opposing stay probabilities on trials following a rare transition. The model-free system increases the value of the first-stage stimulus that was chosen, not the unchosen stimulus, promoting to stay with the current choice. The model-based system promotes switching when receiving a reward after a rare transition, since the alternative first-stage stimulus is more likely to lead to the previously rewarded second-stage state.

To dissociate model-based and model-free control in first-stage choices, I used an one-trial-back stay-switch metric. I examined whether the probability of staying or switching depended on the level of food deprivation (hungry or sated), transition type (common or rare), and reward on previous trial (reward or no reward), using a three-way repeated measures ANOVA (Table 7.1

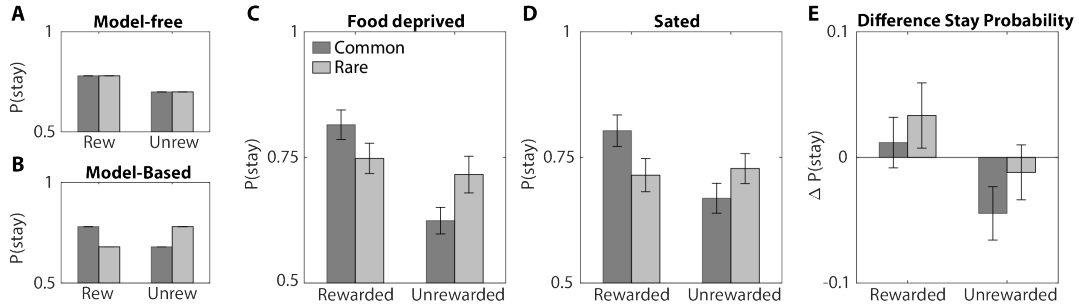


Figure 7.2: **Stay-Switch behaviour at first-stage.** Model-based and model-free strategies for reinforcement learning predict different patterns by which outcomes experienced after the second-stage impact first-stage choices on subsequent trials (Daw et al., 2011). A) Simulation of choice behaviour driven by the model-free system. Rewards increase the likelihood of choosing the same stimulus on the next trial, regardless of whether that reward occurred after a common or rare transition. B) Simulation of choice behaviour driven by the model-based system, which relies on an interaction between transition type and reward outcome. C) Experimental data of food deprived participants. D) Experimental data of sated participants. E) Change in stay probability with food deprivation. Food deprivation increased the tendency to stay after receiving a reward and increased the tendency to switch after reward omission. Error bars represent SEM.

Effect	**F(1,31)**	**p-value**	**Behavioural effect**
Reward	42.39	< 0.0001	Model-free
Transition	0.01	$= 0.921$	
Food deprivation	0.04	$= 0.846$	
Transition × Reward	32.50	< 0.0001	Model-based
Food deprivation × Reward	6.09	$= 0.019$	Food deprivation on model-free
Food deprivation × Transition	2.75	$= 0.107$	
Food deprivation × Transition × Reward	0.07	$= 0.799$	Food deprivation on model-based

Table 7.1: **Statistics stay behaviour: three-way repeated measures ANOVA.**

and Fig. 7.2C–E). Participants used both model-free and model based strategies to solve this task, which is consistent with previous studies (Daw et al., 2011; Deserno et al., 2015; Wunderlich et al., 2012). The key analyses concerned whether food deprivation modulated these learning strategies (Table 7.1). Food deprivation increased the tendency to repeat the same choice after receiving a reward, but not after reward omission. This suggests that food deprived individuals relied more on model-free strategies, but not on model-based strategies. Overall stay behaviour was not affected by food deprivation or an interaction of food deprivation and transition.

The relative contribution of model-free and model-based strategies to choice behaviour can also be assessed with an alternative index (Eq. (7.1) and (7.2)). The current metabolic state modulated the model-free (MF) index [$t_{31} = 2.47$, $p = 0.019$], but not the model-based (MB) index [$t_{31} = 0.26$, $p = 0.799$] (Fig. 7.3A). The session order and the amount of training did not alter MB or MF indices [$p > 0.2$] (Fig. 7.3B). These analyses confirmed that the manipulation of metabolic state, rather than training, caused the effects observed in this study. This observation corresponds with earlier reports that extensive training did not alter the trade-off between the model-based and model-free system (Grosskurth et al., 2019).

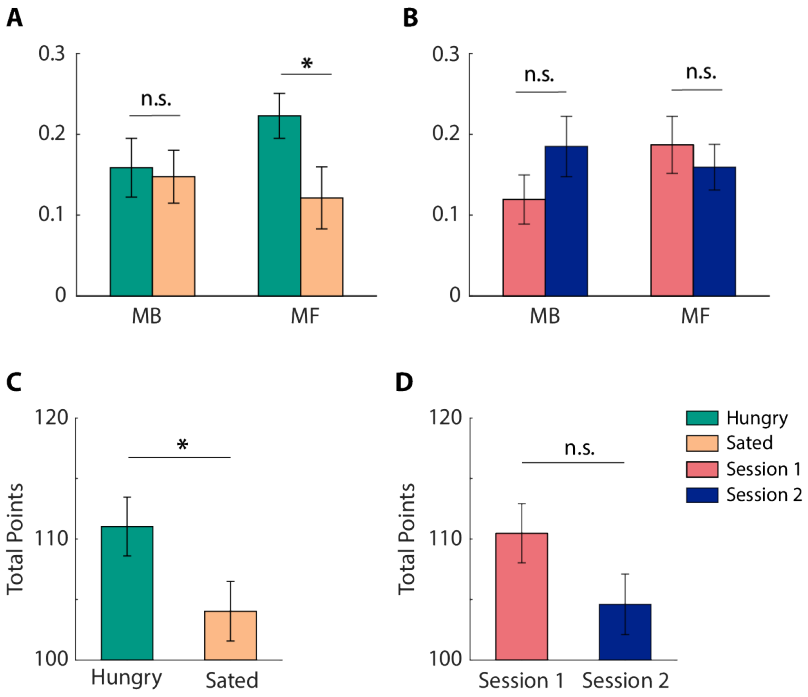


Figure 7.3: **Food deprivation increased model-free control and performance.** A) Food deprivation increased the relative contribution of the model-free (MF), but not model-based (MB), system to stay behaviour at stage 1. B) Training did not significantly alter the contribution of the MB or MF system to choice behaviour. C) Food deprivation increased the number of points earned. D) Session order had no significant impact on the number of points earned. Error bars represent SEM. $^*p < 0.05$.

Food deprivation enhanced performance

Some studies have reported a significant correlation between overall performance and enhanced model-based control (Wunderlich et al., 2012), while others have argued that both strategies are important for optimal performance (Kool et al., 2016). In this study, food deprivation enhanced performance, measured as the number of points earned in the task ($[t_{31} = 2.25, p = 0.032]$; Fig. 7.3C). The total number of points earned on the second testing day was lower, albeit not significant, than on the first testing day ($[t_{31} = 1.84, p = 0.075]$; Fig. 7.3D), suggesting that more experience with the task structure did not improve performance. Although this result may be counter-intuitive, it confirms that food deprivation enhances performance, likely by increasing model-free control.

Other reinforcement-driven behaviours were not affected

In addition to the stay-switch measure reported in Fig. 7.2, there are two other measures of simple reinforcement learning in this task that are independent of previous state transitions. The first measure is stay behaviour for second-stage choices, which is indicative of adaptive behaviour driven by reinforcements. The win-stay and lose-stay probabilities for second-stages were calculated for the two available stimuli together, but separately for the two planets, and divided by the number of wins or losses to give the proportion of stay behaviour for each participant. For this analysis, I assumed that the same state did not have to be visited twice in a row, because this would make the analysis dependent on the first-stage choices. Consequently, I assumed that participants were able to keep an average of the total expected value for a stimulus in the their working memory. Participants were more likely to repeat their previous choice after receiving a reward (main effect of reward $[F_{1,31} = 124.50, p < 0.0001]$). Food deprivation did not alter overall stay behaviour (main effect of

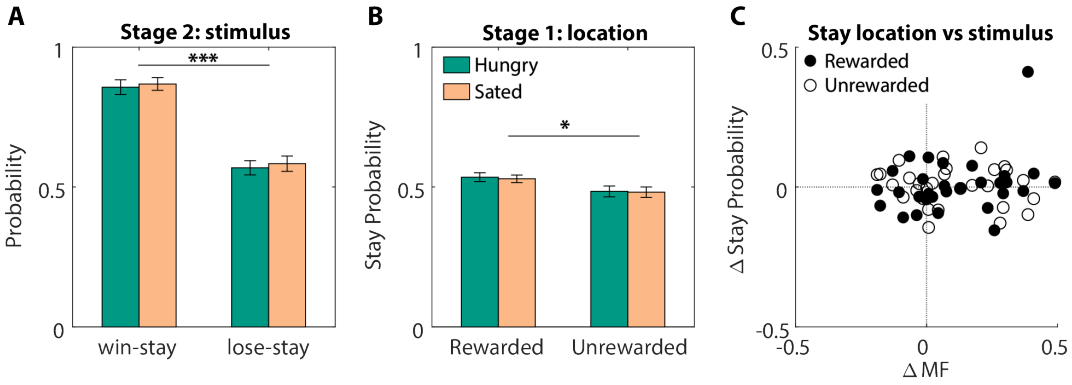


Figure 7.4: **Food deprivation did not alter other types of reinforcement learning.** A) Repetition of second-stage choices was enhanced following rewarded trials and was not altered by food deprivation. B) Repetition of actions was enhanced following rewarded trials, regardless of stimulus identity or the level of food deprivation. C) There was no correlation between first-stage action repetition and the model-free (MF) values of first-stage choices. Action repetition did not contribute to model-free control of stage 1 choices. Error bars represent SEM. * $p < 0.05$, *** $p < 0.001$.

food deprivation [$F_{1,31} < 1$]) or outcome-specific stay behaviour (interaction effect of food deprivation and reward [$F_{1,31} < 1$]; Fig. 7.4A).

The second measure is stay-switch behaviour for outcome-irrelevant information (Shahar et al., 2019). Participants may associate actions (i.e. choosing left or right), rather than stimulus identity, with reward outcomes and repeat an action after receiving a reward on the previous trial for that action. Given that only the stimulus identity, not the stimulus location, has predictive effects, this type of behaviour could be seen as a measure of aberrant incidental learning, in which a value is assigned to actions that were not inherently rewarding. Participants repeated the same action more often after receiving a reward (main effect of reward [$F_{1,31} = 6.34$, $p = 0.017$]). Food deprivation did not alter overall stay behaviour (main effect of food deprivation [$F_{1,31} = 0.19$, $p = 0.664$]) or outcome-specific stay behaviour (interaction effect of food deprivation and reward [$F_{1,31} < 1$]; Fig. 7.4B). The significant effect of reward on action repetition did not affect the contribution of the model-free system to first-stage choices (Fig. 7.4C). Together, these analyses showed that food deprivation selectively affected model-free control of first-stage choices, but not

all types of reinforcement-driven learning.

Computational modelling confirmed model-free effects

Increased model-free control in choice behaviour could be driven by various computational processes that cannot be identified using stay-switch analyses alone. Therefore, I used a computational modelling strategy to assess the effects of food deprivation on behavioural changes and attribute these effects to a specific computational process. The complexity of the task and the contribution of model-based and model-free strategies to choice behaviour were captured by a seven parameter model (Daw et al., 2011).

I used a hierarchical model fitting procedure as described in the Methods section 4.3. The model parameters, denoted by Greek letters, were transformed into a Gaussian scale, denoted by their respective Latin letters, using Eq. (4.7) for the learning rates, the eligibility gain parameter and the weighting parameter, and Eq. (4.9) for the softmax parameters and the perseverance parameter. The statistical significance was tested with respect to the Gaussian scaled model parameters and corrected for multiple comparisons. The quality of the fitting procedure was verified with a parameter recovery analysis (see Methods 4.3). All parameters were well recovered ($0.65 \leq R < 0.92$) and the model fitting procedure did not introduce spurious correlations between the other parameters ($|R| < 0.4$; Fig. 7.5A).

Food deprivation significantly reduced the eligibility gain parameter [λ, $p < 0.001$], suggesting that second stage prediction errors contributed less to the first stage choices. The difference in the second-stage learning rate parameter [α_2, $p = 0.023$] and perseverance parameter [π, $p = 0.017$] was not significant after a Bonferroni correction. There was no significant effect for the first-stage softmax parameter [β_1, $p = 0.084$], second-stage softmax parameter [β_2, $p = 0.508$], first-stage learning rate [α_1, $p = 0.103$], and weighting parameter [ω, $p = 0.800$] (Fig. 7.5B).

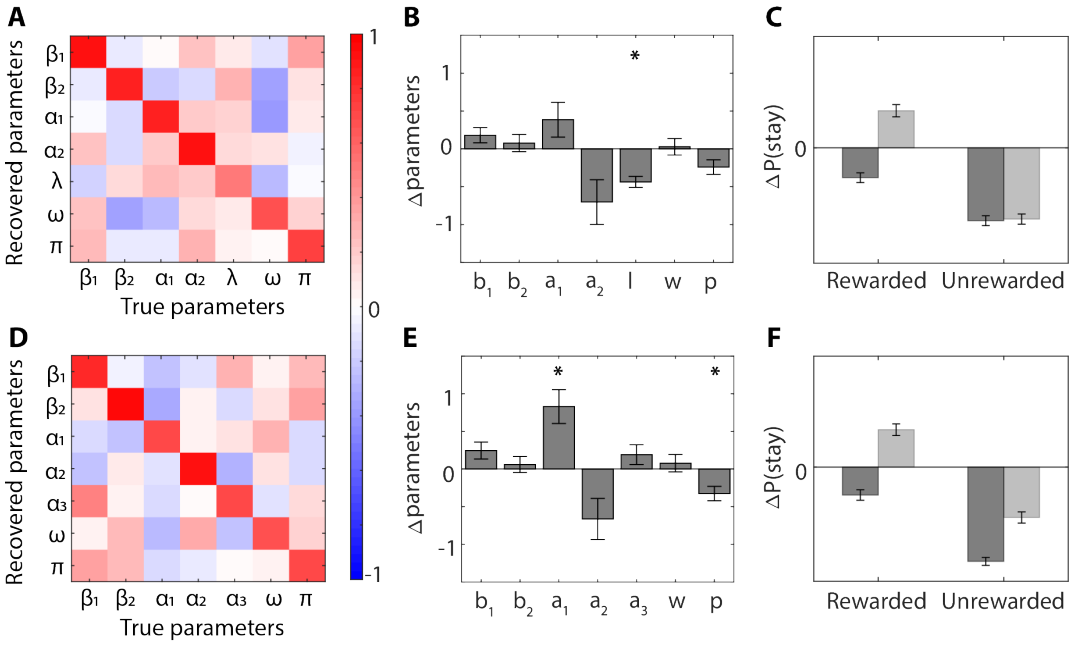


Figure 7.5: **Results Computational Learning Model.** A) Correlation matrix of the free parameters used to generate the simulated data (True parameters) and obtained parameters by applying the parameter estimation procedure on the simulated data (Recovered parameters). Bright red values indicate a strong correlation between the true and recovered parameter value and therefore a good parameter recovery. B) Differences in Gaussian scaled parameters between food deprived and sated conditions in the original model (BIC score: 26112.). Positive values indicate that the parameter estimate was higher when food deprived compared to sated. Statistical significance was tested using paired t-tests on Gaussian scaled parameter estimates. C) Surrogate data simulated with the original model using best fitting parameters for this model (Table 7.2). D) Correlation matrix of the re-parametrised model. E) Differences in parameters estimates for the re-parametrised hybrid model (BIC score: 26100). F) Surrogate data simulated with the re-parametrised model using the best fitting parameters for this model (Table 7.3) revealed a similar pattern of stay behaviour as the experimental data in Fig. 7.2E. Error bars represent SEM. * $p < 0.05/7$ (corrected for multiple comparisons).

The decrease in λ following food deprivation may seem surprising. Eligibility traces are known to contribute to the main effect of reward and high values of λ increase model-free behaviour. The update of model-free values for stage 1 stimuli according to temporal difference learning, Eq. (7.3), relies on the reward prediction error of the state transition and the reward prediction error of stage 2 choices (Fig. 7.6). Second-stage outcomes affect first-stage values by an amount that is scaled by the eligibility trace gain parameter λ and learning

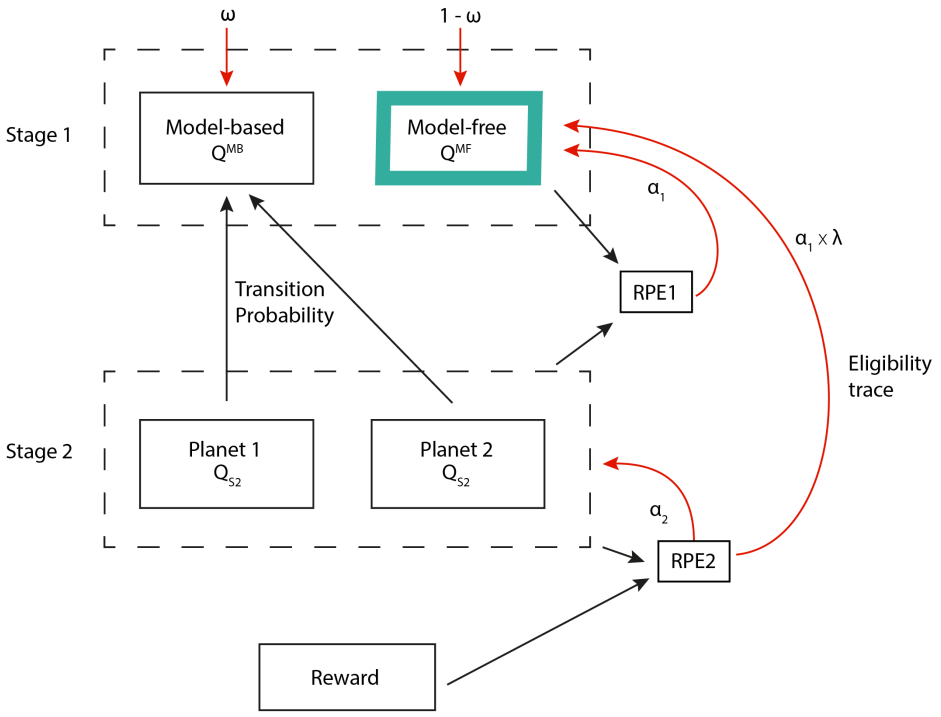


Figure 7.6: **Schematic of value computation.** First-stage choices depend on a model-based and model-free component. Model-based values are computed by multiplying the stimulus with the highest value of both second-stage planets with their respective transition probabilities. Model-free values are updated using a first-stage reward prediction error (RPE1; scaled by α_1), which captures the difference between the value of the chosen option in the first-stage and the chosen option in the second-stage (state transition), and an eligibility trace of the second-stage reward prediction error (RPE2). RPE2 captures the difference between the reward received and the value of the chosen second-stage option (scaled by α_2). The eligibility trace (scaled by $\alpha_1 \times \lambda$) determines the influence RPE2 has on first-stage model-free values. Food deprivation modulated model-free values at stage 1, which is indicated by the green rectangle. Black arrows indicate contribution to. Red arrows indicate that values are scaled by.

rate parameter α_1. This multiplicative relationship may result in a trade-off between λ and α_1. Indeed, Fig. 7.5A shows that food deprivation increased the first-stage learning rate, while it decreased the eligibility gain parameter. When comparing $\alpha_1 \times \lambda$ for hungry with $\alpha_1 \times \lambda$ for sated, the difference in scaling of the eligibility trace was not significant [$p = 0.619$], suggesting that the overall contribution of the eligibility traces was not affected

To formally examine these effects, I re-parametrised the model by replacing $\alpha_1 \times \lambda$ with a new parameter, called α_3. Re-parametrising a model gives

different parameter estimates due to the hierarchical nature of the fitting procedure, which assumes that each parameter has a similar distribution across participants. The model fitting procedure did not introduce spurious correlations ($|R| < 0.4$) and all parameters were recovered well ($0.65 < R < 0.95$; Fig. 7.5D). Using this model, I found that food deprivation significantly increased the learning rate for first-stage choices [α_1, $p < 0.001$], without affecting the contribution of the eligibility trace [α_3, $p = 0.449$], confirming the results above. Food deprivation also significantly reduced the perseverance parameter [π, $p = 0.002$]. The effect of food deprivation on the first-stage softmax parameter [β_1, $p = 0.037$] and second-stage learning rate parameter [α_2, $p = 0.021$] was not significant after Bonferroni correction. I did not find a significant effect for the second-stage softmax parameter [β_2, $p = 0.582$] or weighting parameter [ω, $p = 0.517$] (Fig. 7.5E).

Model comparison (see Methods section 4.3) revealed that the behavioural data was best fitted with the re-parametrised model ($\text{BIC}_{\text{original}} = 26112$ and $\text{BIC}_{\text{re-parametrised}} = 26100$). Surrogate data generated with the best fitted parameters for both models (Table 7.2 & 7.3) confirmed that the difference in stay behaviour (Fig. 7.2E) was better captured with the re-parametrised model (Fig. 7.5F).

	β_1	β_2	α_1	α_2	λ	ω	π
Sated							
25th percentile	3.09	2.13	0.15	0.31	0.59	0.48	0.09
Median	5.59	3.65	0.26	0.48	0.66	0.53	0.13
75th percentile	7.59	5.15	0.47	0.64	0.71	0.62	0.21
Food deprived							
25th percentile	3.50	2.31	0.17	0.10	0.52	0.47	0.07
Median	6.21	4.20	0.39	0.39	0.54	0.54	0.09
75th percentile	9.21	4.93	0.55	0.58	0.58	0.63	0.15

Table 7.2: **Best fitting model parameters estimates obtained with original hybrid model.** Separately shown for sated and food deprived condition as median and quartiles across participants.

	β_1	β_2	α_1	α_2	α_3	ω	π
Sated							
25th percentile	3.04	2.12	0.16	0.31	0.12	0.51	0.10
Median	5.88	3.64	0.25	0.49	0.18	0.58	0.13
75th percentile	7.68	5.15	0.37	0.64	0.24	0.66	0.19
Food deprived							
25th percentile	3.64	2.36	0.32	0.11	0.18	0.47	0.06
Median	5.89	4.24	0.44	0.40	0.21	0.61	0.09
75th percentile	10.81	4.91	0.52	0.58	0.26	0.69	0.16

Table 7.3: **Best fitting model parameters estimates obtained with re-parametrised hybrid model.** Separately shown for sated and food deprived condition as median and quartiles across participants.

7.4 Discussion

Adaptive behaviour requires the ability to respond to changes in reward outcomes and metabolic state to plan actions accordingly. In this study, I used a sequential decision-making task to test whether food deprivation modulates simple reflexive decisions or prospectively planned actions. This study provides evidence that food deprivation enhanced overall performance by increasing model-free control, without changing model-based control. These results suggest that the current metabolic state drives adaptive behaviour for trial-and-error learning, and does not affect cognitive processes that are required for future planning.

Learning strategies

To solve the two-step task there are two dominant strategies that can be employed. People can either use model-free learning and rely on past experiences to update expectations for future outcomes or they can use model-based learning to create an internal model of the task structure based on the transition probabilities and the reward contingencies to maximise their earnings through forward planning. These two strategies predict different patterns by which feedback obtained after the second-stage choice impacts future first-stage choices. If a reward is obtained after the second-stage choice, individuals choose the same

rocket again on the next trial when using a model-free strategy. Model-free strategies ignore the task structure and therefore do not distinguish between common or rare transitions. When using a model-based strategy, individuals decrease their tendency to repeat a first-stage choice after a rare transition that was rewarded, because the alternative first-stage choice is more likely to lead to the previously rewarded second-stage choice. Model-based individuals use the task structure to plan ahead and maximise the likelihood of reaching the alien that was rewarding on the previous trial. In line with previous studies (Daw et al., 2011; Grosskurth et al., 2019; Wunderlich et al., 2012), individuals in the current study used a mixture of model-based and model-free strategies to solve the task.

Hunger increased the importance of model-free control, characterised by a main effect of reward and an increase in the first-stage learning rate parameter derived from computational modelling. Furthermore, food deprived participants obtained significantly more points during the task, indicating that increased model-free control is beneficial to overall performance. This observation may be surprising, because previous studies have reported a positive correlation between model-based control and overall performance (Daw et al., 2011; Wunderlich et al., 2012). The outcome probabilities that drive model-free behaviour change slowly according to Gaussian random walks. To perform optimally, participants continuously have to update the estimates of second-stage outcomes (aliens) according to these outcome probabilities. This implementation ensures that at no point in time a pure model-free strategy is clearly more advantageous to a more complex model-based strategy. The current study showed that hunger did not affect model-based control. The absence of a positive correlation between model-based control and performance may be due to the task design. A recent study demonstrated that the performance on the standard two-step task may not benefit from increased model-based control (Kool et al., 2016). Food deprivation increased performance by increasing the

learning rate to enhance the contribution of model-free behaviour. The additional effects of model-free behaviour on total reward may be small, but suggest that adaptive control of model-free strategies can be advantageous in certain situations, which may be promoted by hunger.

A previous study by Friedel et al. (2014) hinted towards a relationship between model-based control and metabolic need. Friedel et al. indexed the metabolic need of an individual as the change in valuation for food rewards in a separate devaluation paradigm. They observed that people who exhibited greater devaluation also showed more model-based control. In the present study, I did not find a shift in model-based control as a function of metabolic need. By testing the same individuals twice on this sequential decision-making task (once in high metabolic need and once in low metabolic need), I was able to examine the direct effects of changes in metabolic need on the trade-off between model-based and model-free control in the same individual, and correlate this with overall performance. Crucially, my observations and the observations reported by Friedel et al. (2014) are not necessarily incompatible. Changes in metabolic state may not affect model-based behaviour for monetary outcomes, but it could still affect the valuation or devaluation of food rewards. Indeed, participants tested in the current study showed increased valuation for domain-specific food rewards, but not for domain-irrelevant non-food or monetary rewards (Fig. 4.2). Previous studies in animals and humans have shown that the value of a deprived reward is indeed higher in the deprived condition (Aw et al., 2009; Epstein et al., 2003; Pompilio et al., 2006).

Neural substrates of model-free and model-based control

Model-free and model-based learning seem to recruit, at least partially, distinct neural systems. In the sequential decision-making task, model-based prediction error signals have been found in the mPFC and OFC, whereas model-free prediction errors have been found in the VS (Daw et al., 2011; Deserno et al., 2015; Gläscher et al., 2010; Lee et al., 2014; Smittenaar et al., 2013).

Peripheral hormones, including leptin, ghrelin, cholecystokinin, and insulin, modulate goal values assigned to (food) rewards and control associative learning through prediction errors (Higgs et al., 2017). These hormones signal the current level of energy reserves and manipulate behaviour accordingly by interacting with the reward system (Abizaid et al., 2006; Cone et al., 2014; Fulton et al., 2006; Hommel et al., 2006; Kawahara et al., 2009). In particular, ghrelin modulates the earliest stages of reward processing (Naleid et al., 2005), and activates dopaminergic VTA neurons to cause an increase in dopamine release in the NAc and dorsal striatum (Abizaid et al., 2006; Cone et al., 2014; Kawahara et al., 2009). Leptin has the opposite effect on dopamine and reduces food intake (Fulton et al., 2006; Hommel et al., 2006).

Although I do not have an objective read-out of peripheral hormones or dopamine function, I may be able to get an insight into what behaviour is affected in the current task by examining the behavioural results and the results derived from computational modelling.

Model-free learning includes learning from state transitions and from rewards. The computational model used in this study allows for differences in learning or behaviour that may exist between the two stages (Daw et al., 2011). The first-stage learning rate α_1 captures how fast individuals learn from state transitions, and the second-stage learning rate α_2 captures learning from rewards. Hunger increased the tendency to shift behaviour after omission of reward on trials with a common transition, and increased the tendency to stay after receiving a reward on rare trials. This suggests that hunger may have differential effects on learning from state transitions and rewards. Indeed, Fig. 7.5 shows that food deprivation increased α_1, but decreased α_2.

Computational theories map dopaminergic responses onto model-free learning (Montague et al., 1996; Schultz et al., 1997). Given the interaction of peripheral signals with dopaminergic signalling, it is not unsurprising that food deprivation altered model-free control. The basal ganglia plays an important

role in learning and the evaluation of costs and benefits. The value of potential rewards and punishments has been proposed to be mediated by the Go and Nogo neurons, respectively. Elevated levels of tonic dopamine promote learning from positive prediction errors and reduces the effectiveness of negative prediction errors (Frank et al., 2004), thus dopamine could change reward sensitivity. However, computational modelling showed that food deprivation did not alter the second-stage learning rate α_2, suggesting that the reward sensitivity and integration of rewards over multiple trials is not affected. Instead, food deprivation seems to increase learning from state transitions, which may have allowed food deprived individuals to perform better.

Peripheral hormones mostly act on subcortical areas, including the basal ganglia, but food deprivation may also affect cortical areas indirectly. Hunger increases the goal value assigned to food (Papageorgiou et al., 2016), and increases the the hedonic effects and incentive salience of monetary rewards and non-food items (Briers et al., 2006; Xu et al., 2015). In the current study, food deprived participants also earned more points, suggesting that food deprivation also affects domain-irrelevant rewards. It is possible that domain-general effects require the contribution of cortical areas for the explicit valuation and computation of reward that is necessary for learning, motivation and goal-directed choice. Such generalising effects are of particular interest for the consumer market and have broader implications for disorders such as obesity and gambling. Individuals with eating disorders, but not obese individuals, exhibit increased model-free behaviour in the two-step task (Voon et al., 2015). My study shows that even in an unselected population, hunger can increase model-free behaviour.

Limitations

In the current study, I used a Visual Analogue Scale to assess the subjective feeling of hunger. This method has previously been shown to produce reliable measurements for appetite sensations (Flint et al., 2000), but objective and

subjective measures of hunger are not always consistent (Shabat-Simon et al., 2018). This makes it difficult to link changes in subjective sensations of hunger to task performance. In particular, an increase in cognitive demand may reduce feelings of hunger despite an increase in hunger hormones, such as ghrelin (Cummings, 2006). Future studies may want to include objective physiological measures of hunger to capture the variability in hunger levels among individuals (Shabat-Simon et al., 2018). This allows for a systematic study to tease apart the contributions of individual differences to decision-making strategies.

The computational modelling showed that multiple computational processes may be modulated by food deprivation. It is possible that multiple factors, possibly beyond the current model-free/model-based account, affect people's performance on the task. Future studies will be necessary to link the behavioural changes observed in this study to neural activity.

Conclusion

To conclude, I found that increased metabolic need enhanced reflexive model-free control of behaviour, without affecting deliberative model-based control. Although food deprivation has been shown to increase impulsivity, this study provides evidence that relying on a reflexive model-free system does not make hungry people maladaptive, because the deliberate model-based system remains unaffected.

There are no facts, only interpretations.

— Friedrich Nietzsche

8 General discussion

Contents

8.1 Summary 183

8.2 Comparison of experimental and theoretical work . 188

8.3 Considerations and future directions 191

Decision-making relies on adequately integrating internal and external signals to optimise behaviour. In my thesis, I investigated how internal physiological states, such as hunger, affect various aspects of decision-making. I employed an interdisciplinary approach for this investigation. I first developed a mechanistic model to capture the effects of food deprivation on reward valuation, action selection and learning. Then, I designed an experimental study to test the generalising effects of food deprivation on learning and choice for state-irrelevant rewards in human volunteers. In this chapter, I summarise the main findings of each chapter and compare the theoretical predictions with the experimental findings. I also reflect on how these fit with the existing literature and discuss future directions.

8.1 Summary

For the discussion of the results, I distinguish between implicit and explicit decision-making processes, which are thought to involve distinct neural circuits. The dopaminergic BG circuit is thought to mediate implicit, context-dependent action initiation based on the relative probability of positive and negative outcomes. Dopamine activity in the BG determines whether a response is executed or inhibited, according to the subjective value of an action. On the other hand, cortical areas, such as the PFC and OFC, are thought to control explicit action selection based on anticipated rewards because these areas provide estimates of the reinforcement magnitude activated in the working memory. This explicit system estimates the true expected value of reward-related decisions and switches behaviour when reinforcement contingencies change. These two systems have been shown to cooperate and compete for behavioural control (Daw et al., 2005; Frank and O'Reilly, 2006; Ostlund and Balleine, 2005). I discuss all chapters separately and a summary of the findings is presented in Table 8.1.

Theoretical model For the theoretical model, I only considered the involvement of the implicit system. Internal physiological factors, such as hunger, scale the dopaminergic activation and teaching signals. These signals modulate the neural activity and plasticity in the BG, and are therefore likely to alter the motivation for taking actions and learning.

In the model, I defined the motivation to obtain a particular resource as the difference between the desired and the current level of this resource, and proposed how the utility of reinforcements depends on the motivation. I argued that the prediction error encoded in the dopaminergic activity needed to be redefined as the difference between utility and expected utility, which depends on both the objective reinforcement and the motivation. I also demonstrated a possible mechanism by which the evaluation and learning of the utility of actions can be implemented in the basal ganglia network. Go neurons encode

Brain areas / Study	**Basal ganglia circuitry (implicit system)**	**Cortical areas (explicit system)**
Theoretical model	• Internal physiological state modulates the dopamine teaching and activation (motivation) signals. • The utility of an action depends on the reinforcement size and motivation. • High motivation promotes learning from positive consequences, low motivation promotes learning from negative consequences.	
Reward valuation		• Hunger increased the valuation of food rewards, but not of monetary gambles or non-food items.
Approach-avoidance	• Hunger increased the learning rate and caused a trend increase in the sensitivity to positive, confirmatory feedback. There was no change in learned value. Performance was better for positive valence outcomes and low uncertainty pairs (**Training**).	
	• Deprivation level modulated avoidance behaviour, but not approach behaviour (**Test**).	
Risk-taking (implicit)	• Hunger altered learning, particularly in the early stages of the task. There were no differences in learned mean and spread of reward outcomes.	
	• Hunger increased risk-seeking for negative decision contexts, but decreased risk-seeking in positive decision contexts.	
Risk-taking (explicit)		• There was no effect on the explicit evaluation of monetary gambles.
Learning strategies	• Hunger enhanced model-free control.	• There was no effect on model-based control.

Table 8.1: **Summary of experimental and theoretical findings.**

the positive consequences of actions, whereas Nogo neurons encode the negative consequences of actions.

The theory makes strong predictions about the interaction between motivation and reward size. Nogo neurons prevent selecting actions with large reinforcements when the motivation is low, because the desired physiological state will be exceeded. Low motivation emphasises the negative consequences of actions and facilitates Nogo learning, while high motivation emphasises the positive consequences and facilitates Go learning.

For simplicity, I only considered a single physiological dimension (e.g. hunger, salt level) and assumed that each dimension had its own unique values with a maximum desirability. I assumed that this desirability function is concave and non-monotonic. I did not consider resources that are irrelevant to the current physiological state and are stored outside of the body, such as monetary resources. These resources typically have a monotonically increasing diminishing utility for larger reinforcements, because humans always wish to have more money regardless of their current assets. Despite these assets being stored outside the body, the internal physiological state still exerts motivating power over goal-directed actions for monetary resources in a state-dependent manner (Levy et al., 2013; Shabat-Simon et al., 2018; Symmonds et al., 2010). I examined this interaction in more detail in my experimental study.

Reward valuation The subjective value of a reward drives choice behaviour because the option with the highest value is typically chosen to maximise overall gain. In a willingness-to-pay paradigm, I found that hungry participants were indeed willing to pay more for food rewards, but not for non-food items or monetary gambles. This task assessed the explicit valuation of familiar items, without the confounding effect of learning. My findings showed that only the valuation of state-relevant rewards is state-dependent, which suggests that the explicit system may be modulated by food deprivation as well. Previous studies have shown that hungry people show an increased BOLD response in cortical

areas to high calorie food items, whereas sated people show an increased BOLD response to low calorie food items (Siep et al., 2009), which is consistent with my observations. In the theoretical work, I assumed that motivation scales the positive consequences of state-specific rewards, such as food rewards when hungry, but not state-irrelevant rewards. The findings in this experiment support the assumption that reward value is state-dependent, but they also suggest that explicit valuation may already occur at the level of the cortex.

Approach versus avoidance behaviour The theoretical work makes strong predictions about approach and avoidance behaviour for food rewards. It predicts that food deprivation increases the level of dopamine and promotes approach behaviour, whereas satiety reduces the level of dopamine and promotes avoidance behaviour. I tested whether food deprivation also modulated approach and avoidance behaviour for monetary outcomes. In particular, I examined the effects of food deprivation on learning and choice behaviour for probabilistic positive and negative outcomes. I investigated whether the biasing effects on behaviour were the result of changes in sensitivity to reward prediction errors or reward valence. I found that food deprivation altered choice behaviour, but not learning. The valence and probability of outcomes affected the performance during the learning phase, but not the learned values at the end of training. Food deprivation increased the speed of learning, but not the overall accuracy. Food deprivation differentially modulated approach and avoidance behaviour in the gain and loss domain. Hungry participants performed equally well for avoiding or approaching outcomes with positive or negative valence, whereas sated individuals showed a pronounced difference. In particular, sated individuals were better at avoiding infrequent gains compared to avoiding frequent losses. The outcome probabilities for the gain and loss domain were statistically identical and should not have induced any biases. These findings suggest that the context or framing of outcomes, gains versus losses, also affects choice behaviour and learning. It also suggests that once the

valence of the symbols is learned, individuals reframe their expectations such that neutral outcomes become punishing in the gain condition and rewarding in the loss condition.

Risk-taking Dopamine plays an important role in risk-seeking propensities in healthy individuals (Rigoli et al., 2016) and PD patients (Gallagher et al., 2007). Food deprivation enhances dopamine release (Aitken et al., 2016; Cone et al., 2016; Papageorgiou et al., 2016) and increases risk-taking behaviour (Levy et al., 2013; Shabat-Simon et al., 2018; Symmonds et al., 2010), which suggests that a causal relationship between food deprivation, dopamine levels and risk-taking might exist. Following the predictions of the payoff-cost model (Mikhael and Bogacz, 2016), I hypothesised that learned risks may be mediated by values encoded in the Go and Nogo pathways of the basal ganglia. Hunger may enhance risk-seeking by increasing the dopaminergic activation signals that controls the tendency to take risks. I tested whether food deprivation altered risk preferences for monetary outcomes in various decision contexts in which the risk was either learned or described. For learned risks, I found that food deprivation increased risk-taking for negatively framed choices, and decreased risk-taking for positively framed choices. This suggests that individuals make choices with respect to the immediate context (i.e. which of the presented options is better) and the long-term memory (i.e. how good is this option relative to all the options experienced in the past). Food deprivation did not modulate risk preferences in described risks, suggesting that cortical processes involved in the evaluation explicit risk-taking may not be influenced by hunger.

Learning strategies For the theoretical work, I assumed that individuals reliably infer their current energy reserves and plan their actions in such a way to maintain or restore a physiological balance by creating an internal model of the body in the world. Long-term nutritional planning may rely on both reactive and prospective mechanisms. Previous research has shown that food

deprivation may affect cognitive control (Higgs et al., 2017) and increases impulsivity and reward sensitivity (Bartholdy et al., 2016; Skrynka and Vincent, 2019). I tested whether food deprivation promoted simple reflexive decisions or prospectively-planned actions, when both strategies can be employed simultaneously. I found that food deprivation increased reflexive model-free control of behaviour, but not deliberative model-based control. The implicit reward system is thought to be involved in model-free control, whereas the explicit reward system is involved in model-based control. These findings suggest that food deprivation modulates the implicit system to enhance the reliance on "gut-instinct", without impairing cognitive control.

8.2 Comparison of experimental and theoretical work

The theoretical work is based on the assumption that hunger increases the dopaminergic activation and teaching signals. Hunger could enhance learning of outcome contingencies, by increasing the dopaminergic teaching signal. Hunger could energise all actions, not only the ones specific to the motivational drive, by enhancing the dopaminergic activation signal. The theoretical work only makes predictions about learning and action selection processes that are mediated by the basal ganglia circuitry (Table 8.2). Of all the experimental tasks presented in this thesis, the model makes the strongest predictions about approach and avoidance behaviour (chapter 5), and implicit risk-taking (chapter 6). The model predicts that the tendency to consume large portions of food or other reinforcers is reduced by low dopamine levels, as seen in satiety, while it is increased by high dopamine levels, as seen in hunger. Therefore, I expected that hungry participants would be better at approach behaviour, while sated participants would be better at avoidance behaviour. Additionally, when facing uncertain outcomes, the payoff-cost model predicts that risky actions have higher Go weights relative to safe actions with low uncertainty. Consequently,

Predictions / Study	**Hunger increases dopamine motivation signal**	**Hunger increases learning from rewards**
Approach-avoidance	• **Prediction:** Hunger increases approach behaviour, while satiety increases avoidance behaviour (**Test**). • **Data:** Hunger mainly affected avoidance behaviour. Hunger tended to alter avoidance behaviour differently depending on the valence of outcomes (**Test**).	• **Prediction:** Hunger increases learning from positive feedback (**Training**). • **Data:** Hunger increased learning and tended to increase the sensitivity to positive, confirmatory feedback. Hunger did not affect the learned value (**Training**).
Risk-taking (implicit)	• **Prediction:** Hunger increases overall risk-seeking behaviour • **Data:** Hunger did not alter overall risk-seeking, but modulated risk-taking in a decision context dependent manner.	• **Prediction:** Increase in learning rate (for state-relevant rewards) • **Data:** Hunger decreased the learning rate for the spread in reward, but tended to increase learning of average reward.

Table 8.2: **Comparison of theoretical predictions and empirical findings**

an increased dopamine level promotes the selection of a risky action over a safe action. Therefore, I expected that participants would become more risk-seeking for decisions involving risk, independently of the decision context.

The experimental findings did not fully support the theoretical predictions. Food deprivation did not promote approach behaviour or increase overall risk-taking (Table 8.2). Interestingly, these two experimental studies show similarities in their design. First, the outcome probabilities need to be learned in both tasks. Although, the implicit risk-task did not have a separate training phase, only the second half of the trials were used to measure risk-taking behaviour at which point learning had stabilised (Fig. 6.5). Second, both tasks involve positive and negative decision contexts. Positive decision contexts concerned choices between outcomes that either led to an outcome with positive valence or an outcome that was positive relative to the average reward in the task. The opposite was true for negative decision contexts.

The experimental results of the two studies also show similarities. Food

deprivation tended to increase learning in both tasks. Food deprivation had opposite effects in positive and negative decision contexts. In the approach-avoidance task, sated participants had stronger biases for avoidance behaviour. Compared to hungry individuals, they were better at avoiding the least rewarding option in the gain domain and were worse at avoiding the least rewarding option in the loss domain. In the implicit risk task, sated individuals were risk-averse for negatively framed choices, but risk-seeking for positively framed choices. Hungry participants showed risk-neutral behaviour for both decision contexts. These findings suggest that the framing of choices is important for choice behaviour. Sated individuals exhibit strong biases in choice behaviour that depend on the context and valence of outcomes. Hungry individuals show minimal biases for avoidance behaviour and base their choices on the expected value of uncertain prospects without overweighting or underweighting uncertain outcomes. Hunger has been previously associated with maladaptive behaviour, however the results in this study show that hunger makes people more "rational" in their behaviour, which may positively contribute to performance.

Most of my experimental findings suggest an involvement of the implicit decision-making system in the state-dependent modulation of learning and action selection. However, the experimental findings do not fully correspond to the predictions I made based on the effect food deprivation has on dopamine function. This might be because: 1) hunger does not modulate the dopaminergic activation signal to energise all actions, or 2) the evaluation of domain-irrelevant rewards rely on neural circuits that extend beyond the basal ganglia.

Existing literature has shown a causal link between food deprivation, dopamine levels and state-dependent goal-directed behaviour in rodents (Aitken et al., 2016; Papageorgiou et al., 2016). Evidence for the effect of food deprivation on dopamine in human volunteers is limited. Some research suggests that the evaluation of food and monetary rewards may recruit different brain areas. In an incentive delay task, reward related anticipation for food reward, but not

monetary rewards, increased BOLD responses in the VS. During reward receipt, an interaction between deprivation level and reward size was measured in the medial OFC for monetary outcomes, but not food outcomes (Simon et al., 2017). This study provides valuable insight into how food and monetary rewards are evaluated under different deprivation levels. Future research will have to address the same processes in tasks that require goal-directed actions.

Existing studies implicate brain areas that extend beyond the basal ganglia in the evaluation of outcomes. Emerging evidence suggest that positive and negative consequences are also encoded by different populations of amygdala neurons (Namburi et al., 2015). An up-shift in homeostatic need, caused by acute food deprivation, promotes the circuitry that encodes positive consequences, whereas a down-shift by satiety promotes the circuitry that encodes negative consequences (Calhoon et al., 2018). Furthermore, different cortical regions preferentially project to either Go or Nogo neurons, suggesting that positive and negative consequences of actions are already differentially encoded in the cortex (Wall et al., 2013). Finally, the insular cortex encodes reward uncertainty for explicit risks (Preuschoff et al., 2008). However, the differential effects of food deprivation on implicit and explicit risk-taking suggest that the reward uncertainty for implicit risks may be encoded elsewhere. Contextual information from the cortex and valence coding from the amygdala may exert control over the Go and Nogo pathway to select or inhibit actions based on past experiences in similar contexts.

8.3 Considerations and future directions

Neural circuitries involved in feeding are more complex than considered in the theory. In particular, I did not consider the hypothalamus, which drives feeding behaviour in various parallel and redundant circuitries (Betley et al., 2013). Additionally, state-dependent reward valuation may already occur at the level of the cortex. Contextual inputs, risk prediction errors and valence coding may

rely on other brain areas, such as the insula and amygdala. Thus, the current model may be oversimplified and will need to be extended to account for neural processing outside the BG and incorporate new empirical findings.

Future studies will be necessary to systematically investigate the contribution of these candidate areas in the decision-making processes that are modulated by food deprivation. Imaging studies in animals could increase our understanding of the contribution of specific neural populations, such as the contribution of Go and Nogo neurons in action selection as suggested in chapter 3. Additional imaging and stimulation studies may reveal whether specific cortical areas or the amygdala indeed contribute to learning and action selection under different deprivation levels. It will also be important to differentiate studies investigating acute food deprivation from studies that rely on chronic deprivation. Chronic food deprivation is associated with adverse effects, such as stress, and induce long-term changes in plasticity and gene expression, which also impact the level of dopamine (Pothos et al., 1995a,b). Imaging studies in human volunteers could advance our understanding of decision-making processes that include the interaction between internal and external signals, especially if they are combined with physiological read-outs of biological markers or hormones. In particular, studies that compare goal-directed actions for food and monetary outcomes will help contextualise the current studies.

Bibliography

Aberg, K. C., Doell, K. C., and Schwartz, S. (2016). Linking Individual Learning Styles to Approach-Avoidance Motivational Traits and Computational Aspects of Reinforcement Learning. *PloS ONE*, 11(11):e0166675.

Aberman, J. E., Ward, S. J., and Salamone, J. D. (1998). Effects of dopamine antagonists and accumbens dopamine depletions on time-constrained progressive-ratio performance. *Pharmacology Biochemistry and Behavior*, 61(4):341–348.

Abizaid, A., Liu, Z.-W., Andrews, Z. B., Shanabrough, M., Borok, E., Elsworth, J. D., Roth, R. H., Sleeman, M. W., Picciotto, M. R., Tschop, M. H., Gao, X.-B., and Horvath, T. L. (2006). Ghrelin modulates the activity and synaptic input organization of midbrain dopamine neurons while promoting appetite. *The Journal of clinical investigation*, 116(12):3229–3239.

Abler, B., Walter, H., Erk, S., Kammerer, H., and Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *NeuroImage*, 31(2):790–795.

Aitken, T. J., Greenfield, V. Y., and Wassum, K. M. (2016). Nucleus accumbens core dopamine signaling tracks the need-based motivational value of food-paired cues. *Journal of neurochemistry*, 136(5):1026–1036.

Aldrich, J. (1997). R. A. Fisher and the making of maximum likelihood 1912 - 1922. *Statistical Science*, 12(3):162–176.

Alhadeff, A. L., Rupprecht, L. E., and Hayes, M. R. (2012). GLP-1 neurons in the nucleus of the solitary tract project directly to the ventral tegmental area and nucleus accumbens to control for food intake. *Endocrinology*, 153(2):647–658.

Anand, B. and Brobeck, J. (1951). Hypothalamic control of food intake in rats and cats. *The Yale journal of biology and medicine*, 24(2):123–140.

Anderberg, R. H., Hansson, C., Fenander, M., Richard, J. E., Dickson, S. L., Nissbrandt, H., Bergquist, F., and Skibicka, K. P. (2016). The Stomach-Derived Hormone Ghrelin Increases Impulsive Behavior. *Neuropsychopharmacology*, 41:1199–1209.

Anderson, R. I., Bush, P. C., and Spear, L. P. (2013). Environmental manipulations alter age differences in attribution of incentive salience to reward-paired cues. *Behavioural Brain Research*, 257(607):83–89.

Andrews, Z. B., Erion, D., Beiler, R., Liu, Z. W., Abizaid, A., Zigman, J., Elsworth, J. D., Savitt, J. M., DiMarchi, R., Tschoep, M., Roth, R. H., Gao, X. B., and Horvath, T. L. (2009). Ghrelin promotes and protects nigrostriatal dopamine function via a UCP2-dependent mitochondrial mechanism. *Journal of Neuroscience*, 29(45):14057–14065.

Aponte, Y., Atasoy, D., and Sternson, S. M. (2011). AGRP neurons are sufficient to orchestrate feeding behavior rapidly and without training. *Nature Neuroscience*, 14(3):351–355.

Aw, J. M., Holbrook, R., Burt de Perera, T., and Kacelnik, A. (2009). State-dependent valuation learning in fish: Banded tetras prefer stimuli associated with greater past deprivation. *Behavioural Processes*, 81(2):333–336.

Balleine, B. (1994). Asymmetrical Interactions between Thirst and Hunger in Pavlovian-instrumental Transfer. *The Quarterly Journal of Experimental Psychology Section B*, 47(2):211–231.

Balleine, B. W. (1992). Instrumental performance following a shift in primary motivation depends on incentive learning. *Journal of experimental psychology. Animal behavior processes*, 18(3):236–250.

Barrett, L. F. and Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, 16(7):419–429.

Bartholdy, S., Cheng, J., Schmidt, U., Campbell, I. C., and O'Daly, O. G. (2016). Task-based and questionnaire measures of inhibitory control are differentially affected by acute food restriction and by motivationally salient food stimuli in healthy adults. *Frontiers in Psychology*, 7(AUG):1–13.

Bateson, M. and Kacelnik, A. (1996). Rate currencies and the foraging starling: the fallacy of the averages revisited. *Behavioral Ecology*, 7(3):341–352.

Bautista, L. M., Tinbergen, J., and Kacelnik, A. (2001). To walk or to fly? How birds choose among foraging modes. *Proceedings of the National Academy of Sciences of the United States of America*, 98(3):1089–1094.

Bayer, H. M. and Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1):129–141.

Becker, G. M., Degroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226 – 232.

Berghorst, L. H., Bogdan, R., Frank, M. J., and Pizzagalli, D. A. (2013). Acute stress selectively reduces reward sensitivity. *Frontiers in human neuroscience*, 7:133.

Berke, J. D. (2018). What does dopamine mean? *Nature Neuroscience*, 21:787–793.

Bernsmeier, C., Weisskopf, D. M., Pflueger, M. O., Mosimann, J., Campana, B., Terracciano, L., Beglinger, C., Heim, M. H., and Cajochen, C. (2015). Sleep Disruption and Daytime Sleepiness Correlating with Disease Severity and Insulin Resistance in Non-Alcoholic Fatty Liver Disease: A Comparison with Healthy Controls. *PLoS ONE*, 10(11):e0143293.

Berridge, K. C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology*, 191(3):391–431.

Berridge, K. C. (2012). From prediction error to incentive salience: mesolimbic computation of reward motivation. *The European journal of neuroscience*, 35(7):1124–1143.

Berridge, K. C. and Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, 28(3):309–369.

Berridge, K. C. and Schulkin, J. (1989). Palatability shift of a salt-associated incentive during sodium depletion. *The Quarterly Journal of Experimental Psychology*, 41(2):121–138.

Betley, J. N., Cao, Z. F. H., Ritola, K. D., and Sternson, S. M. (2013). Parallel, redundant circuit organization for homeostatic control of feeding behavior. *Cell*, 155(6):1337–1350.

Betley, J. N., Xu, S., Cao, Z. F. H., Gong, R., Magnus, C. J., Yu, Y., and Sternson, S. M. (2015). Neurons for hunger and thirst transmit a negative-valence teaching signal. *Nature*, 521(7551):180–185.

Billes, S. K., Simonds, S. E., and Cowley, M. A. (2012). Leptin reduces food intake via a dopamine D2 receptor-dependent mechanism. *Molecular Metabolism*, 1(1-2):86–93.

Bindra, D. (1978). How adaptive behavior is produced: A perceptual-motivational alternative to response reinforcements. *Behavioral and Brain Sciences*, 1(1):41–52.

Bond, A. and Lader, M. (1974). The use of analogue scales in rating subjective feelings. *British Journal of Medical Psychology*, 47(3):211–218.

Boorman, E. D. and Sallet, J. (2009). Mean-variance or prospect theory? The nature of value representations in the human brain. *Journal of Neuroscience*, 29(25):7945–7947.

Briers, B., Pandelaere, M., Dewitte, S., and Warlop, L. (2006). Hungry for Money. *Psychological Science*, 17(11):939–943.

Brito e Abreu, F. and Kacelnik, A. (1999). Energy budgets and risk-sensitive foraging in starlings. *Behavioral Ecology*, 10(3):338–345.

Burgess, C. R., Ramesh, R. N., Sugden, A. U., Levandowski, K. M., Minnig, M. A., Fenselau, H., Lowell, B. B., and Andermann, M. L. (2016). Hunger-Dependent Enhancement of Food Cue Responses in Mouse Postrhinal Cortex and Lateral Amygdala. *Neuron*, 91(5):1154–1169.

Burton, A. L., Hay, P., Kleitman, S., Smith, E., Raman, J., Swinbourne, J., Touyz, S. W., and Abbott, M. J. (2017). Confirmatory factor analysis and examination of the psychometric properties of the eating beliefs questionnaire. *BMC Psychiatry*, 17(237).

Bush, R. R. and Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, 58(6):413–423.

Bushong, B., King, L. M., Camerer, C. F., and Rangel, A. (2010). Pavlovian Processes in Consumer Choice: The Physical Presence of a Good Increases Willingness-to-Pay. *American Economic Review*, 100:1556–1571.

Cai, X., Kim, S., and Lee, D. (2011). Heterogeneous Coding of Temporally Discounted Values in the Dorsal and Ventral Striatum during Intertemporal Choice. *Neuron*, 69(1):170–182.

Calhoon, G. G., Sutton, A. K., Chang, C.-J., Libster, A. M., Glober, G. F., Leveque, C. L., Murphy, G. D., Namburi, P., Leppla, C. A., Siciliano, C. A., Wildes, C. P., Kimchi, E. Y., Beyeler, A., and Tye, K. M. (2018). Acute Food Deprivation Rapidly Modifies Valence-Coding Microcircuits in the Amygdala. *bioRxiv*, page 285189.

Camerer, C. and Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, 5(4):325–370.

Camilleri, A. R. and Newell, B. R. (2011). When and why rare events are underweighted: A direct comparison of the sampling, partial feedback, full feedback and description choice paradigms. *Psychonomic Bulletin and Review*, 18(2):377–384.

Canessa, N., Crespi, C., Motterlini, M., Baud-Bovy, G., Chierchia, G., Pantaleo, G., Tettamanti, M., and Cappa, S. F. (2013). The functional and structural neural basis of individual differences in loss aversion. *Journal of Neuroscience*, 33(36):14307–14317.

Caraco, T. (1981). Energy budgets, risk and foraging preferences in dark-eyed juncos (Junco hyemalis). *Behavioral Ecology and Sociobiology*, 8(3):213–217.

Caraco, T. and Lima, S. L. (1985). Foraging juncos: interaction of reward mean and variability. *Animal Behaviour*, 33(1):216–224.

Carlson, J. N., Herrick, K. F., Baird, J. L., and Glick, S. D. (1987). Selective enhancement of dopamine utilization in the rat prefrontal cortex by food deprivation. *Brain Research*, 400(1):200–203.

Carpenter, R. H. S. (2004). Homeostasis: a plea for a unified approach. *Advances in Physiology Education*, 28(4):180–187.

Carr, K. D., Tsimberg, Y., Berman, Y., and Yamamoto, N. (2003). Evidence of increased dopamine receptor signaling in food-restricted rats. *Neuroscience*, 119(4):1157–1167.

Carver, C. and White, T. (1994). Bahavioural inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales. *Journal of personality and social psychology*, 67(2):319–333.

Cavanagh, J. F., Frank, M. J., and Allen, J. J. (2011). Social stress reactivity alters reward and punishment learning. *Social Cognitive and Affective Neuroscience*, 6(3):311–320.

Chen, P. S., Yang, Y. K., Yeh, T. L., Lee, I. H., Yao, W. J., Chiu, N. T., and Lu, R. B. (2008). Correlation between body mass index and striatal dopamine transporter availability in healthy volunteers-A SPECT study. *NeuroImage*, 40(1):275–279.

Chew, B., Hauser, T. U., Papoutsi, M., Magerkurth, J., Dolan, R. J., and Rutledge, R. B. (2019). Endogenous fluctuations in the dopaminergic midbrain drive behavioral choice variability. *Proceedings of the National Academy of Sciences*, 116(37):18732–18737.

Chib, V. S., Rangel, A., Shimojo, S., and O'Doherty, J. P. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *Journal of Neuroscience*, 29(39):12315–12320.

Christopoulos, G. I., Tobler, P. N., Bossaerts, P., Dolan, R. J., and Schultz, W. (2009). Neural correlates of value, risk, and risk aversion contributing to decision making under risk. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(40):12574–83.

Clark, L., Bechara, A., Damasio, H., Aitken, M. R. F., Sahakian, B. J., Robbins, T. W., and Clark, L. (2008). Differential effects of insular and ventromedial prefrontal cortex lesions on risky decision-making. *Brain*, 131:1311–1322.

Collins, A. G. E. and Frank, M. J. (2014). Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review*, 121(3):337–366.

Compan, V., Walsh, B. T., Kaye, W., and Geliebter, A. (2015). How does the brain implement adaptive decision making to eat? *Journal of Neuroscience*, 35(41):13868–13878.

Cone, J. J., Fortin, S. M., McHenry, J. A., Stuber, G. D., McCutcheon, J. E., and Roitman, M. F. (2016). Physiological state gates acquisition and expression of mesolimbic reward prediction signals. *Proceedings of the National Academy of Sciences of the United States of America*, 113(7):1943–1948.

Cone, J. J., McCutcheon, J. E., and Roitman, M. F. (2014). Ghrelin acts as an interface between physiological state and phasic dopamine signaling. *The Journal of Neuroscience*, 34(14):4905–4913.

Cools, R., Frank, M. J., Gibbs, S. E., Miyakawa, A., Jagust, W., and D'Esposito, M. (2009). Striatal dopamine predicts outcome-specific reversal learning and its sensitivity to dopaminergic drug administration. *Journal of Neuroscience*, 29(5):1538–1543.

Corbit, L. H., Leung, B. K., and Balleine, B. W. (2013). The role of the amygdala-striatal pathway in the acquisition and performance of goal-directed instrumental actions. *Journal of Neuroscience*, 33(45):17682–17690.

Cox, S. M., Frank, M. J., Larcher, K., Fellows, L. K., Clark, C. A., Leyton, M., and Dagher, A. (2015). Striatal D1 and D2 signaling differentially predict learning from positive and negative outcomes. *NeuroImage*, 109:95–101.

Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A., and Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, 7(2):189–195.

Cummings, D. E. (2006). Ghrelin and the short- and long-term regulation of appetite and body weight. *Physiology and Behavior*, 89(1):71–84.

D'Acremont, M., Lu, Z.-L., Li, X., Van der Linden, M., and Bechara, A. (2009). Neural correlates of risk prediction error during reinforcement learning in humans. *NeuroImage*, 47(4):1929–1939.

D'Agostino, G., Lyons, D. J., Cristiano, C., Burke, L. K., Madara, J. C., Campbell, J. N., Garcia, A. P., Land, B. B., Lowell, B. B., Dileone, R. J., and Heisler, L. K. (2016). Appetite controlled by a cholecystokinin nucleus of the solitary tract to hypothalamus neurocircuit. *eLife*, 5(MARCH2016).

Davidson, T. L., Altizer, A. M., Benoit, S. C., Walls, E. K., and Powley, T. L. (1997). Encoding and selective activation of 'metabolic memories' in the rat. *Behavioral Neuroscience*, 111(5):1014–1030.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, 69(6):1204–1215.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711.

Day, J. J., Jones, J. L., Wightman, R. M., and Carelli, R. M. (2010). Phasic nucleus accumbens dopamine release encodes effort- and delay-related costs. *Biological psychiatry*, 68(3):306–9.

Day, J. J., Roitman, M. F., Wightman, R. M., and Carelli, R. M. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nature Neuroscience*, 10(8):1020–1028.

De Martino, B., Kumaran, D., Seymour, B., and Dolan, R. J. (2006). Frames, biases and rational decision-making in the human brain. *Science*, 313(5787):684–687.

de Ridder, D., Kroese, F., Adriaanse, M., and Evers, C. (2014). Always Gamble on an Empty Stomach: Hunger Is Associated with Advantageous Decision Making. *PLoS ONE*, 9(10):e111081.

Deroche, V., Marinelli, M., Maccari, S., Le Moal, M., Simon, H., and Piazza, P. V. (1995). Stress-induced sensitization and glucocorticoids. I. Sensitization of dopamine-dependent locomotor effects of amphetamine and morphine depends on stress-induced corticosterone secretion. *Journal of Neuroscience*, 15(11):7181–7188.

Derryberry, D. and Reed, M. A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of Abnormal Psychology*, 111(2):225–236.

Deserno, L., Huys, Q. J., Boehme, R., Buchert, R., Heinze, H. J., Grace, A. A., Dolan, R. J., Heinz, A., and Schlagenhauf, F. (2015). Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences of the United States of America*, 112(5):1595–1600.

Dickinson, A. and Balleine, B. (1990). Motivational control of instrumental performance following a shift from thirst to hunger. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology*, 42B(4):413–431.

Dickinson, A. and Balleine, B. W. (2002). The role of learning in the operation of motivational systems. In *Stevens' Handbook of Experimental Psychology*, pages 497–534. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Doll, B. B., Bath, K. G., Daw, N. D., and Frank, M. J. (2016). Variability in dopamine genes dissociates model-based and model-free reinforcement learning. *Journal of Neuroscience*, 36(4):1211–1222.

Domingos, A. I., Vaynshteyn, J., Voss, H. U., Ren, X., Gradinaru, V., Zang, F., Deisseroth, K., de Araujo, I. E., and Friedman, J. (2011). Leptin regulates the reward value of nutrient. *Nature Neuroscience*, 14(12):1562 – 1589.

Dorris, M. C. and Glimcher, P. W. (2004). Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron*, 44(2):365–378.

Dreyer, J. K., Herrik, K. F., Berg, R. W., and Hounsgaard, J. D. (2010). Influence of phasic and tonic dopamine release on receptor activation. *Journal of Neuroscience*, 30(42):14273–14283.

Dunovan, K., Vich, C., Clapp, M., Verstynen, T., and Rubin, J. (2019). Reward-driven changes in striatal pathway competition shape evidence evaluation in decision-making. *PLOS Computational Biology*, 15(5):e1006998.

Elias, C. F., Saper, C. B., MaratosFlier, E., Tritos, N. A., Lee, C., Kelly, J., Tatro, J. B., Hoffman, G. E., Ollmann, M. M., Barsh, G. S., Sakurai, T., Yanagisawa, M., and Elmquist, J. K. (1998). Chemically defined projections linking the mediobasal hypothalamus and the lateral hypothalamic area. *Journal of Comparative Neurology*, 402(4):442–459.

Elliott, R., Rees, G., and Dolan, R. J. (1999). Ventromedial prefrontal cortex mediates guessing. *Neuropsychologia*, 37(4):403–411.

Elmquist, J. K., Bjørbæk, C., Ahima, R. S., Flier, J. S., and Saper, C. B. (1998). Distributions of leptin receptor mRNA isoforms in the rat brain. *Journal of Comparative Neurology*, 395(4):535–547.

Epstein, L. H., Truesdale, R., Wojcik, A., Paluch, R. A., and Raynor, H. A. (2003). Effects of deprivation on hedonics and reinforcing value of food. *Physiology and Behavior*, 78(2):221–227.

Feher da Silva, C. and Hare, T. A. (2018). A note on the analysis of two-stage task results: How changes in task structure affect what model-free and model-based strategies predict about the effects of reward and transition on the stay probability. *PLoS ONE*, 13(4):e0195328.

Figlewicz, D. P., Bennett, J. L., Naleid, A. M. D., Davis, C., and Grimm, J. W. (2006). Intraventricular insulin and leptin decrease sucrose self-administration in rats. *Physiology and Behavior*, 89(4):611–616.

Figlewicz, D. P., Evans, S. B., Murphy, J., Hoen, M., and Baskin, D. G. (2003). Expression of receptors for insulin and leptin in the ventral tegmental area/substantia nigra (VTA/SN) of the rat. *Brain Research*, 964(1):107–115.

Figlewicz, D. P., Sipols, A. J., Bennett, J., Evans, S. B., Kaiyala, K., and Benoit, S. C. (2004). Intraventricular insulin and leptin reverse place preference conditioned with high-fat diet in rats. *Behavioral Neuroscience*, 118(3):479–487.

Fiorillo, C. D., Newsome, W. T., and Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. *Nature Neuroscience*, 11(8):966–973.

Fiorillo, C. D., Tobler, P. N., and Schultz, W. (2003). Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. *Science*, 299(5614):1898 – 1902.

Fitzgerald, T. H., Seymour, B., Bach, D. R., and Dolan, R. J. (2010). Differentiable neural substrates for learned and described value and risk. *Current Biology*, 20(20):1823–1829.

Flint, A., Raben, A., Blundell, J. E., and Astrup, A. (2000). Reproducibility, power and validity of visual analogue scales in assessment of appetite sensations in single test meal studies. *International Journal of Obesity*, 24(1):38–48.

Fontanesi, L., Gluth, S., Spektor, M. S., and Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin and Review*, 26(4):1099–1121.

Franchina, J. J. and Brown, T. S. (1971). Reward magnitude shift effects in rats with hippocampal lesions. *Journal of comparative and physiological psychology*, 76:365–370.

Frank, M. J. (2005). Dynamic Dopamine Modulation in the Basal Ganglia: A Neurocomputational Account of Cognitive Deficits in Medicated and Nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience*, 17(1):51–72.

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., and Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, 104(41):16311 – 16316.

Frank, M. J. and O'Reilly, R. C. (2006). A mechanistic account of striatal dopamine function in human cognition: Psychopharmacological studies with cabergoline and haloperidol. *Behavioral Neuroscience*, 120(3):497–517.

Frank, M. J., Seeberger, L. C., and O'reilly, R. C. (2004). By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science*, 306(5703):1940 – 1943.

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*, 40(2):13–86.

Friedel, E., Koch, S. P., Wendt, J., Heinz, A., Deserno, L., and Schlagenhauf, F. (2014). Devaluation and sequential decisions: linking goal-directed and model-based behavior. *Frontiers in Human Neuroscience*, 8:587.

Friedman, M. and Savage, L. J. (1952). The Expected-Utility Hypothesis and the Measurability of Utility. *Journal of Political Economy*, 60(6):463 – 474.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

Fulton, S., Pissios, P., Manchon, R. P., Stiles, L., Frank, L., Pothos, E. N., Maratos-Flier, E., and Flier, J. S. (2006). Leptin Regulation of the Mesoaccumbens Dopamine Pathway. *Neuron*, 51(6):811–822.

Gallagher, D. A., O'Sullivan, S. S., Evans, A. H., Lees, A. J., and Schrag, A. (2007). Pathological gambling in Parkinson's disease: Risk factors and differences from dopamine dysregulation. An analysis of published case series. *Movement Disorders*, 22(12):1757–1763.

Gan, J. O., Walton, M. E., and Phillips, P. E. M. (2009). Dissociable cost and benefit encoding of future rewards by mesolimbic dopamine. *Nature Neuroscience*, 13(1):25.

Gawley, D. J., Timberlake, W., and Lucas, G. A. (1988). Anticipatory drinking in rats: Compensatory adjustments in the local rate of intake. *Physiology and Behavior*, 42(3):297–302.

Gerfen, C. R. and Surmeier, D. J. (2011). Modulation of striatal projection systems by dopamine. *Annual review of neuroscience*, 34:441–66.

Gilovich, T., Griffin, D. W., and Kahneman, D. (2002). *Heuristics and biases*. New York: Cambridge University Press.

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–595.

Gong, R., Xu, S., Hermundstad, A., Yu, Y., and Sternson, S. M. (2020). Hindbrain Double-Negative Feedback Mediates Palatability-Guided Food and Water Consumption. *Cell*, 182:1–17.

Gremel, C. M. and Lovinger, D. M. (2017). Associative and sensorimotor cortico-basal ganglia circuit roles in effects of abused drugs. *Genes, Brain and Behavior*, 16(1):71–85.

Grogan, J. P., Tsivos, D., Smith, L., Knight, B. E., Bogacz, R., Whone, A., and Coulthard, E. J. (2017). Effects of dopamine on reinforcement learning and consolidation in Parkinson's disease. *eLife*, 6:e26801.

Grosskurth, E. D., Bach, D. R., Economides, M., Huys, Q. J. M., and Holper, L. (2019). No substantial change in the balance between model-free and model-based control via training on the two-step task. *PLoS Computational Biology*, 15(11):e1007443.

Guitart-Masip, M., Fuentemilla, L., Bach, D. R., Huys, Q. J. M., Dayan, P., Dolan, R. J., and Duzel, E. (2011). Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(21):7867–75.

Gunaydin, L. A. and Kreitzer, A. C. (2016). CorticoBasal Ganglia Circuit Function in Psychiatric Disease. *Annual Review of Physiology*, 78(1):327–350.

Gurney, K., Prescott, T. J., and Redgrave, P. (2001). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics*, 84(6):401–410.

Haber, S. N., Fudge, J. L., and McFarland, N. R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *Journal of Neuroscience*, 20(6):2369–2382.

Hamdi, A., Porter, J., and Chandan, P. (1992). Decreased striatal D2 dopamine receptors in obese Zucker rats: changes during aging. *Brain Research*, 589(2):338–340.

Han, J. E., Frasnelli, J., Zeighami, Y., Malik, S., Jones-Gotman, M., and Dagher Correspondence, A. (2018). Ghrelin Enhances Food Odor Conditioning in Healthy Humans: An fMRI Study. *Cell Reports*, 25:2643–2652.

Hartmann, M. N., Hager, O. M., Tobler, P. N., and Kaiser, S. (2013). Parabolic discounting of monetary rewards by physical effort. *Behavioural Processes*, 100:192–196.

Hassenstab, J. J., Sweet, L. H., Del Parigi, A., McCaffery, J. M., Haley, A. P., Demos, K. E., Cohen, R. A., and Wing, R. R. (2012). Cortical thickness of the cognitive control network in obesity and successful weight loss maintenance: A preliminary MRI study. *Psychiatry Research - Neuroimaging*, 202(1):77–79.

Helson, H. (1964). *Adaptation-level theory: an experimental and systematic approach to behavior.* New York, Harper and Row.

Hernández-López, S., Tkatch, T., Perez-Garci, E., Galarraga, E., Bargas, J., Hamm, H., and Surmeier, D. J. (2000). D2 Dopamine Receptors in Striatal Medium Spiny Neurons Reduce L-Type Ca2+ Currents and Excitability via a Novel PLCβ1IP3Calcineurin-Signaling Cascade. *Journal of Neuroscience*, 20(24):8987–8995.

Hertwig, R. (2012). The psychology and rationality of decisions from experience. *Synthese*, 187(1):269–292.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions From Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, 15(8):534–539.

Hertwig, R. and Erev, I. (2009). The descriptionexperience gap in risky choice. *Trends in Cognitive Sciences*, 13(12):517–523.

Higgs, S., Spetter, M. S., Thomas, J. M., Rotshtein, P., Lee, M., Hallschmid, M., and Dourish, C. T. (2017). Interactions between metabolic, reward and cognitive processes in appetite control: Implications for novel weight management therapies. *Journal of Psychopharmacology*, 31(11):1460–1474.

Hoebel, B. G. and Teitelbaum, P. (1962). Hypothalamic control of feeding and self-stimulation. *Science*, 135(3501):375–377.

Hommel, J. D., Trinko, R., Sears, R. M., Georgescu, D., Liu, Z. W., Gao, X. B., Thurmon, J. J., Marinelli, M., and DiLeone, R. J. (2006). Leptin Receptor Signaling in Midbrain Dopamine Neurons Regulates Feeding. *Neuron*, 51(6):801–810.

Houston, A. I. (1991). Risk-sensitive foraging theory and operant psychology. *Journal of the Experimental Analysis of Behaviour*, 56(3):585–589.

Howe, M. W., Tierney, P. L., Sandberg, S. G., Phillips, P. E. M., and Graybiel, A. M. (2013). Prolonged Dopamine Signalling in Striatum Signals Proximity and Value of Distant Rewards. *Nature*, 500(7464):575–579.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., and Camerer, C. F. (2005). Neuroscience: Neural systems responding to degrees of uncertainty in human decision-making. *Science*, 310(5754):1680–1683.

Hsu, M., Krajbich, I., Zhao, C., and Camerer, C. F. (2009). Neural response to reward anticipation under risk is nonlinear in probabilities. *Journal of Neuroscience*, 29(7):2231–2237.

Huang, C. C., Chou, D., Yeh, C. M., and Hsu, K. S. (2016). Acute food deprivation enhances fear extinction but inhibits long-term depression in the lateral amygdala via ghrelin signaling. *Neuropharmacology*, 101:36–45.

Huettel, S. A., Song, A. W., and McCarthy, G. (2005). Decisions under uncertainty: Probabilistic context influences activation of prefrontal and parietal cortices. *Journal of Neuroscience*, 25(13):3304–3311.

Huettel, S. A., Stowe, C. J., Gordon, E. M., Warner, B. T., and Platt, M. L. (2006). Neural Signatures of Economic Preferences for Risk and Ambiguity. *Neuron*, 49(5):765–775.

Hull, C. L. (1943). *Principles of Behaviour: an introduction to behavior theory.* Appleton-Century-Crofts, Inc, New York.

Hunger, L., Kumar, A., and Schmidt, R. (2020). Abundance compensates kinetics: Similar effect of dopamine signals on D1 and D2 receptor populations. *The Journal of Neuroscience*, 40(14):2868–2881.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., and Roiser, J. P. (2012). Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, 8(3).

Iodice, P., Ferrante, C., Brunetti, L., Cabib, S., Protasi, F., Walton, M. E., and Pezzulo, G. (2017). Fatigue modulates dopamine availability and promotes flexible choice reversals during decision making. *Scientific Reports*, 7(1):535.

Jennings, J. H., Rizzi, G., Stamatakis, A. M., Ung, R. L., and Stuber, G. D. (2013). The inhibitory circuit architecture of the lateral hypothalamus orchestrates feeding. *Science*, 341(6153):1517–1521.

Jennings, J. H., Ung, R. L., Resendez, S. L., Stamatakis, A. M., Taylor, J. G., Huang, J., Veleta, K., Kantak, P. A., Aita, M., Shilling-Scrivo, K., Ramakrishnan, C., Deisseroth, K., Otte, S., and Stuber, G. D. (2015). Visualizing hypothalamic network dynamics for appetitive and consummatory behaviors. *Cell*, 160(3):516–527.

Jessup, R. K. and O'Doherty, J. P. (2010). Decision Neuroscience: Choices ofDescription and of Experience. *Current Biology*, 20(20):R881–R883.

Jewett, D. C., Cleary, J., Levine, A. S., Schaal, D. W., and Thompson, T. (1995). Effects of neuropeptide Y, insulin, 2-deoxyglucose, and food deprivation on food-motivated behavior. *Psychopharmacology*, 120(3):267–271.

Jocham, G., Klein, T. A., and Ullsperger, M. (2011). Dopamine-Mediated Reinforcement Learning Signals in the Striatum and Ventromedial Prefrontal Cortex Underlie Value-Based Choices. *Journal of Neuroscience*, 31(5):1606–1613.

Joel, D. and Weiner, I. (2000). The connections of the dopaminergic system with the striatum in rats and primates: An analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience*, 96(3):451–474.

Kable, J. W. and Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, 10(12):1625–1633.

Kacelnik, A. and Bateson, M. (1996). Risky theories - The effects of variance on foraging decisions. *American Zoologist*, 36(4):402–434.

Kacelnik, A. and Bateson, M. (1997). Risk-sensitivity: crossroads for theories of decision-making. *Trends in Cognitive Sciences*, 1(8):304–309.

Kacelnik, A. and Marsh, B. (2002). Cost can increase preference in starlings. *Animal Behaviour*, 63(2):245–250.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: an Analyses of Decision under Risk. *Econometrica*, 47(2):263–191.

Kawagoe, R., Takikawa, Y., and Hikosaka, O. (1998). Expectation of reward modulates cognitive signals in the basal ganglia. *Nature Neuroscience*, 1(5):411–416.

Kawahara, Y., Kawahara, H., Kaneko, F., Yamada, M., Nishi, Y., Tanaka, E., and Nishi, A. (2009). Peripherally administered ghrelin induces bimodal effects on the mesolimbic dopamine system depending on food-consumptive states. *Neuroscience*, 161(3):855–864.

Kelley, A. E., Baldo, B. A., Pratt, W. E., and Will, M. J. (2005). Corticostriatal-hypothalamic circuitry and food motivation: Integration of energy, action and reward. *Physiology and Behavior*, 86(5):773–795.

Keramati, M. and Gutkin, B. (2014). Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife*, 3:e04811.

Kim, H., Sul, J. H., Huh, N., Lee, D., and Jung, M. W. (2009). Role of striatum in updating values of chosen actions. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(47):14701–12.

Kirk, J. and Logue, A. (1997). Effects of Deprivation Level on Humans' Self-Control for Food Reinforcers. *Appetite*, 28(3):215–226.

Klein, T. A., Neumann, J., Reuter, M., Hennig, J., Von Cramon, D. Y., and Ullsperger, M. (2007). Genetically determined differences in learning from errors. *Science*, 318(5856):1642–1645.

Klein-Flügge, M. C., Kennerley, S. W., Saraiva, A. C., Penny, W. D., and Bestmann, S. (2015). Behavioral modeling of human choices reveals dissociable effects of physical effort and temporal delay on reward devaluation. *PLOS Computational Biology*, 11(3):e1004116.

Knoch, D., Gianotti, L. R., Pascual-Leone, A., Treyer, V., Regard, M., Hohmann, M., and Brugger, P. (2006). Disruption of right prefrontal cortex by low-frequency repetitive transcranial magnetic stimulation induces risk-taking behavior. *Journal of Neuroscience*, 26(24):6469–6472.

Knutson, B., Adams, C. M., Fong, G. W., Hommer, D., Varner, J., Kerich, M., Lionetti, T., and Walker, J. (2001). Anticipation of Increasing Monetary Reward Selectively Recruits Nucleus Accumbens. *The Journal of Neuroscience*, 21:RC159.

Knutson, B. and Bossaerts, P. (2007). Neural antecedents of financial decisions. *Journal of Neuroscience*, 27(31):8174–8177.

Kool, W., Cushman, F. A., and Gershman, S. J. (2016). When Does Model-Based Control Pay Off? *PLOS Computational Biology*, 12(8):e1005090.

Krashes, M. J., Koda, S., Ye, C. P., Rogan, S. C., Adams, A. C., Cusher, D. S., Maratos-Flier, E., Roth, B. L., and Lowell, B. B. (2011). Rapid, reversible activation of AgRP neurons drives feeding behavior in mice. *Journal of Clinical Investigation*, 121(4):1424–1428.

Kravitz, A. V., Freeze, B. S., Parker, P. R., Kay, K., Thwin, M. T., Deisseroth, K., and Kreitzer, A. C. (2010). Regulation of parkinsonian motor behaviors by optogenetic control of basal ganglia circuitry. *Nature*, 466(7306):622.

Krieckhaus, E. E. (1970). "Innate recognition" aids rats in sodium regulation. *Journal of Comparative and Physiological Psychology*, 73(1):117–122.

Krieckhaus, E. E. and Wolf, G. (1968). Acquisition of Sodium By Rats: Interaction of Innate Mechanisms and Latent Learning. *Journal of Comparative and Physiological Psychology*, 65(2):197–201.

Kroemer, N. B., Lee, Y., Pooseh, S., Eppinger, B., Goschke, T., and Smolka, M. N. (2019). L-DOPA reduces model-free control of behavior by attenuating the transfer of value to action. *NeuroImage*, 186:113–125.

Krügel, U., Schraft, T., Kittner, H., Kiess, W., and Illes, P. (2003). Basal and feeding-evoked dopamine release in the rat nucleus accumbens is depressed by leptin. *European Journal of Pharmacology*, 482(1-3):185–187.

Kruschwitz, J. D., Kausch, A., Brovkin, A., Keshmirian, A., Paulus, M. P., Goschke, T., and Walter, H. (2019). Self-control is linked to interoceptive inference: Craving regulation and the prediction of aversive interoceptive states induced with inspiratory breathing load. *Cognition*, 193:104028.

Kuhnen, C. M. and Knutson, B. (2005). The neural basis of financial risk taking. *Neuron*, 47(5):763–770.

Labouèbe, G., Liu, S., Dias, C., Zou, H., Wong, J. C., Karunakaran, S., Clee, S. M., Phillips, A. G., Boutrel, B., and Borgland, S. L. (2013). Insulin induces long-term depression of ventral tegmental area dopamine neurons via endocannabinoids. *Nature Neuroscience*, 16(3):300–308.

Lau, B. and Glimcher, P. W. (2008). Value Representations in the Primate Striatum duringMatching Behavior. *Neuron*, 58(3):451–463.

Lee, H. J., Weitz, A. J., Bernal-Casas, D., Duffy, B. A., Choy, M. K., Kravitz, A. V., Kreitzer, A. C., and Lee, J. H. (2016). Activation of Direct and Indirect Pathway Medium Spiny Neurons Drives Distinct Brain-wide Responses. *Neuron*, 91(2):412–424.

Lee, S. W., Shimojo, S., and O'Doherty, J. P. (2014). Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. *Neuron*, 81(3):687–699.

Levy, D. J., Thavikulwat, A. C., and Glimcher, P. W. (2013). State Dependent Valuation: The Effect of Deprivation on Risk Preferences. *PLoS ONE*, 8(1):e53978.

Levy, I., Snell, J., Nelson, A. J., Rustichini, A., and Glimcher, P. W. (2010). Neural Representation of Subjective Value Under Risk and Ambiguity. *Journal of Neurophysiology*, 103(2):1036–1047.

Lighthall, N. R., Gorlick, M. A., Schoeke, A., Frank, M. J., and Mather, M. (2013). Stress modulates reinforcement learning in younger and older adults. *Psychology and aging*, 28(1):35–46.

Livneh, Y., Ramesh, R. N., Burgess, C. R., Levandowski, K. M., Madara, J. C., Fenselau, H., Goldey, G. J., Diaz, V. E., Jikomes, N., Resch, J. M., Lowell, B. B., and Andermann, M. L. (2017). Homeostatic circuits selectively gate food cue responses in insular cortex. *Nature*, 546(7660):611–616.

Livneh, Y., Sugden, A. U., Madara, J. C., Essner, R. A., Flores, V. I., Sugden, L. A., Resch, J. M., Lowell, B. B., and Andermann, M. L. (2020). Estimation of Current and Future Physiological States in Insular Cortex. *Neuron*, 105(6):1094–1111.e10.

Lopez, M., Balleine, B. W., and Dickinson, A. (1992). Incentive learning and the motivational control of instrumental performance by thirst. *Animal Learning & Behavior*, 20(4):322–328.

Louie, K., Glimcher, P. W., and Webb, R. (2015). Adaptive neural coding: from biological to behavioral decision-making. *Current opinion in behavioral sciences*, 5:91–99.

Ludvig, E. A., Madan, C. R., and Spetch, M. L. (2014). Extreme Outcomes Sway Risky Decisions from Experience. *Journal of Behavioral Decision Making*, 27:146–156.

Ludvig, E. A. and Spetch, M. L. (2011). Of Black Swans and Tossed Coins: Is the Description-Experience Gap in Risky Choice Limited to Rare Events? *PLoS ONE*, 6(6):e20262.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*, volume 13. Cambridge University Press, Cambrigde, 7.2 edition.

Madan, C. R., Ludvig, E. A., and Spetch, M. L. (2014). Remembering the best and worst of times: Memories for extreme outcomes bias risky decisions. *Psychonomic Bulletin and Review*, 21(3):629–636.

Madan, C. R., Spetch, M. L., and Ludvig, E. A. (2015). Rapid makes risky: Time pressure increases risk seeking in decisions from experience. *Journal of Cognitive Neuroscience*, 27(8):921–928.

Mai, B., Sommer, S., and Hauber, W. (2012). Motivational states influence effort-based decision making in rats: The role of dopamine in the nucleus accumbens. *Cognitive, Affective, & Behavioral Neuroscience*, 12(1):74–84.

Malik, S., McGlone, F., Bedrossian, D., and Dagher, A. (2008). Ghrelin Modulates Brain Activity in Areas that Control Appetitive Behavior. *Cell Metabolism*, 7(5):400–409.

Malvaez, M., Greenfield, V. Y., Wang, A. S., Yorita, A. M., Feng, L., Linker, K. E., Monbouquette, H. G., and Wassum, K. M. (2015). Basolateral amygdala rapid glutamate release encodes an outcome-specific representation vital for reward-predictive cues to selectively invigorate reward-seeking actions. *Scientific Reports*, 5:12511.

Marchand, W. R. (2010). Cortico-basal ganglia circuitry: A review of key research and implications for functional connectivity studies of mood and anxiety disorders. *Brain Structure and Function*, 215(2):73–96.

Marcott, P. F., Mamaligas, A. A., and Ford, C. P. (2014). Phasic dopamine release drives rapid activation of striatal D2-receptors. *Neuron*, 84(1):164–176.

Markowitz, H. (1952). The Utility of Wealth. *Journal of Political Economy*, 60:151–158.

Marsh, B., Schuck-Paim, C., and Kacelnik, A. (2004). Energetic state during learning affects foraging choices in starlings. *Behavioral Ecology*, 15(3):396–399.

McClure, S. M., Daw, N. D., and Read Montague, P. (2003). A computational substrate for incentive salience. *Opinion TRENDS in Neurosciences*, 26(8):423 – 428.

McClure, S. M., Laibson, D. I., Loewenstein, G., Cohen, J. D., and Camerer, C. F. (2004). Separate Neural Systems Value Immediate and Delayed Monetary Rewards. *Science*, 306(5695):503–507.

McCoy, A. N. and Platt, M. L. (2005). Risk-sensitive neurons in macaque posterior cingulate cortex. *Nature Neuroscience*, 8(9):1220–1227.

Mebel, D. M., Wong, J. C., Dong, Y. J., and Borgland, S. L. (2012). Insulin in the ventral tegmental area reduces hedonic feeding and suppresses dopamine concentration via increased reuptake. *European Journal of Neuroscience*, 36(3):2336–2346.

Medic, N., Ziauddeen, H., Vestergaard, M. D., Henning, E., Schultz, W., Farooqi, I. S., Fletcher, P. C., Sadaf Farooqi, I., Fletcher, P. C., Paul, X., and Fletcher, C. (2014). Dopamine Modulates the Neural Representation of Subjective Value of Food in Hungry Subjects. *The Journal of Neuroscience*, 34(50):16856–16864.

Middleton, F. A. and Strick, P. L. (2000). Basal ganglia and cerebellar loops: Motor and cognitive circuits. *Brain Research Reviews*, 31:236–250.

Mikhael, J. G. and Bogacz, R. (2016). Learning Reward Uncertainty in the Basal Ganglia. *PLOS Computational Biology*, 12(9):e1005062.

Mirenowicz, J. and Schultz, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, 379(6564):449–451.

Mishra, S., Gregson, M., and Lalumì Ere, M. L. (2012). Framing effects and risk-sensitive decision making. *British Journal of Psychology*, 103:83–97.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill Inc., New York, NY, USA, 1 edition.

Mohebi, A., Pettibone, J. R., Hamid, A. A., Wong, J.-M. T., Vinson, L. T., Patriarchi, T., Tian, L., Kennedy, R. T., and Berke, J. D. (2019). Dissociable dopamine dynamics for learning and motivation. *Nature*, 570(7759):65–70.

Mollenauer, S. O. (1971). Shifts in deprivation level: Different effects depending on amount of preshift training. *Learning and Motivation*, 2(1):58–66.

Möller, M. and Bogacz, R. (2019). Learning the payoffs and costs of actions. *PLOS Computational Biology*, 15(2):e1006285.

Möller, M., Grohn, J., Manohar, S., and Bogacz, R. (2020). Reward prediction errors induce risk-seeking. *bioRxiv*, page 067751.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of neuroscience*, 16(5):1936–1947.

Moscarello, J. M., Ben-Shahar, O., and Ettenberg, A. (2007). Dynamic interaction between medial prefrontal cortex and nucleus accumbens as a function of both motivational state and reinforcer magnitude: A c-Fos immunocytochemistry study. *Brain Research*, 1169(1):69–76.

Moscarello, J. M., Ben-Shahar, O., and Ettenberg, A. (2009). Effects of food deprivation on goal-directed behavior, spontaneous locomotion, and c-Fos immunoreactivity in the amygdala. *Behavioural Brain Research*, 197(1):9–15.

Moscarello, J. M., Ben-Shahar, O., and Ettenberg, A. (2010). External incentives and internal states guide goal-directed behavior via the differential recruitment of the nucleus accumbens and the medial prefrontal cortex. *Neuroscience*, 170(2):468–477.

Moyer, J. T., Wolf, J. A., and Finkel, L. H. (2007). Effects of Dopaminergic Modulation on the Integrative Properties of the Ventral Striatal Medium Spiny Neuron. *Journal of Neurophysiology*, 98(6):3731–3748.

Murray, E. A. and Rudebeck, P. H. (2013). The drive to strive: Goal generation based on current needs. *Frontiers in Neuroscience*, 7:112.

Naleid, A. M., Grace, M. K., Cummings, D. E., and Levine, A. S. (2005). Ghrelin induces feeding in the mesolimbic reward pathway between the ventral tegmental area and the nucleus accumbens. *Peptides*, 26(11):2274–2279.

Namburi, P., Beyeler, A., Yorozu, S., Calhoon, G. G., Halbert, S. A., Wichmann, R., Holden, S. S., Mertens, K. L., Anahtar, M., Felix-Ortiz, A. C., Wickersham, I. R., Gray, J. M., and Tye, K. M. (2015). A circuit mechanism for differentiating positive and negative associations. *Nature*, 520(7549):675–678.

Nederkoorn, C., Guerrieri, R., Havermans, R. C., Roefs, A., and Jansen, A. (2009). The interactive effect of hunger and impulsivity on food intake and purchase in a virtual supermarket. *International Journal of Obesity*, 33(8):905–912.

Niv, Y., Daw, N. D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, 191(3):507–520.

Niv, Y., Edlund, J. A., Dayan, P., and O'Doherty, J. P. (2012). Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. *Journal of Neuroscience*, 32(2):551–562.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science*, 304(5669):452–454.

O'Doherty, J., Rolls, E. T., Francis, S., Bowtell, R., McGlone, F., Kobal, G., Renner, B., and Ahne, G. (2000). Sensory-specific satiety-related olfactory activation of the human orbitofrontal cortex. *NeuroReport*, 11(4):893–897.

Olds, J. and Milner, P. (1954). Positive reinforcement produced by electrical stimulation septal and other regions of rat brain. *Journal of Comparative and Physiological Psychology*, 47(6):419–427.

O'Neill, M. and Schultz, W. (2010). Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. *Neuron*, 68(4):789–800.

Orsini, C. A., Moorman, D. E., Young, J. W., Setlow, B., and Floresco, S. B. (2015). Neural mechanisms regulating different forms of risk-related decision-making: Insights from animal models. *Neuroscience & Biobehavioral Reviews*, 58:147–167.

Ostlund, S. B. and Balleine, B. W. (2005). Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. *Journal of Neuroscience*, 25(34):7763–7770.

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., and Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, 110(52):20941–20946.

Padoa-Schioppa, C. and Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, 441(7090):223–226.

Palminteri, S., Khamassi, M., Joffily, M., and Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, 6(1):8096.

Palminteri, S., Lefebvre, G., Kilford, E. J., and Blakemore, S. J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS computational biology*, 13(8):e1005684.

Papageorgiou, G. K., Baudonnat, M., Cucca, F., and Walton, M. E. (2016). Mesolimbic Dopamine Encodes Prediction Errors in a State-Dependent Manner. *Cell Reports*, 15(2):221–228.

Park, H., Lee, D., and Chey, J. (2017). Stress enhances model-free reinforcement learning only after negative outcome. *PLoS ONE*, 12(7):e0180588.

Parkes, S. L. and Balleine, B. W. (2013). Incentive memory: Evidence the basolateral amygdala encodes and the insular cortex retrieves outcome values to guide choice between goal-directed actions. *Journal of Neuroscience*, 33(20):8753–8763.

Paulus, M. P. (2007). Decision-making dysfunctions in psychiatry - Altered homeostatic processing? *Science*, 318(5850):602–606.

Paulus, M. P., Rogalsky, C., Simmons, A., Feinstein, J. S., and Stein, M. B. (2003). Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. *NeuroImage*, 19(4):1439–1448.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., and Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106):1042–1045.

Pfaffly, J., Michaelides, M., Wang, G.-J., Pessin, J. E., Volkow, N. D., and Thanos, P. K. (2010). Leptin increases striatal dopamine D2 receptor binding in leptin-deficient obese (ob/ob) mice. *Synapse (New York, N.Y.)*, 64(7):503–10.

Plassmann, H., O'Doherty, J., and Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *Journal of Neuroscience*, 27(37):9984–9988.

Plassmann, H., O'Doherty, J. P., and Rangel, A. (2010). Appetitive and aversive goal values are encoded in the medial orbitofrontal cortex at the time of decision making. *Journal of Neuroscience*, 30(32):10799–10808.

Platt, M. L. and Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, 400(6741):233–238.

Platt, M. L. and Huettel, S. A. (2008). Risky business: the neuroeconomics of decision making under uncertainty. *Nature neuroscience*, 11(4):398–403.

Pompilio, L. and Kacelnik, A. (2005). State-dependent learning and suboptimal choice: when starlings prefer long over short delays to food. *Animal Behaviour*, 70(3):571–578.

Pompilio, L., Kacelnik, A., and Behmer, S. T. (2006). State-dependent learned valuation drives choice in an invertebrate. *Science*, 311(5767):1613–1615.

Pothos, E. N., Creese, I., and Hoebel, B. G. (1995a). Restricted eating with weight loss selectively decreases extracellular dopamine in the nucleus accumbens and alters dopamine response to amphetamine, morphine, and food intake. *Journal of Neuroscience*, 15(10):6640–6650.

Pothos, E. N., Hernandez, L., and Hoebel, B. G. (1995b). Chronic Food Deprivation Decreases Extracellular Dopamine in the Nucleus Accumbens: Implications for a Possible Neurochemical Link Between Weight Loss and Drug Abuse. *Obesity Research*, 3(S4):525S–529S.

Preuschoff, K. and Bossaerts, P. (2007). Adding Prediction Risk to the Theory of Reward Learning. *Annals of the New York Academy of Sciences*, 1104(1):135–146.

Preuschoff, K., Bossaerts, P., and Quartz, S. R. (2006). Neural Differentiation of Expected Reward and Risk in Human Subcortical Structures. *Neuron*, 51(3):381–390.

Preuschoff, K., Quartz, S. R., and Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, 28(11):2745–2752.

Preuschoff, K., 't Hart, B. M., and Einhäuser, W. (2011). Pupil Dilation Signals Surprise: Evidence for Noradrenaline's Role in Decision Making. *Frontiers in neuroscience*, 5:115.

Rangel, A. (2013). Regulation of dietary choice by the decision-making circuitry. *Nature Neuroscience*, 16(12):1717–1724.

Rangel, A., Camerer, C., and Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature reviews. Neuroscience*, 9(7):545–556.

Reppucci, C. J. and Petrovich, G. D. (2012). Learned food-cue stimulates persistent feeding in sated rats. *Appetite*, 59(2):437–447.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory*, 21(6):64–99.

Rigoli, F., Rutledge, R. B., Dayan, P., and Dolan, R. J. (2016). The influence of contextual reward statistics on risk preference. *NeuroImage*, 128:74–84.

Robbins, T. W. and Everitt, B. J. (2007). A role for mesencephalic dopamine in activation: commentary on Berridge (2006). *Psychopharmacology*, 191(3):433–437.

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., and Yiend, J. (1997). 'Oops!' Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*, 35(6):747–758.

Roesch, M. R., Calu, D. J., and Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience*, 10(12):1615–24.

Rogers, R. D., Tunbridge, E. M., Bhagwagar, Z., Drevets, W. C., Sahakian, B. J., and Carter, C. S. (2003). Tryptophan depletion alters the decision-making of healthy volunteers through altered processing of reward cues. *Neuropsychopharmacology*, 28(1):153–162.

Rothschild, M. and Stiglitz, J. E. (1970). Increasing risk: I. A definition. *Journal of Economic Theory*, 2(3):225–243.

Rudebeck, P. H., Behrens, T. E., Kennerley, S. W., Baxter, M. G., Buckley, M. J., Walton, M. E., and Rushworth, M. F. (2008). Frontal cortex subregions play distinct roles in choices between actions and stimuli. *Journal of Neuroscience*, 28(51):13775–13785.

Rummery, G. A. and Niranjan, M. (1994). On-Line Q-Learning Using Connectionist Systems. Technical report, Cambridge University Engineering Department.

Rushworth, M. F. and Behrens, T. E. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, 11(4):389–397.

Salamone, J. D., Steinpreis, R. E., McCullough, L. D., Smith, P., Grebel, D., and Mahan, K. (1991). Haloperidol and nucleus accumbens dopamine depletion suppress lever pressing for food but increase free food consumption in a novel food choice procedure. *Psychopharmacology*, 104(4):515–521.

Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of Action-Specific Reward Values in the Striatum. *Science*, 310(5752):1337 LP – 1340.

Schultz, W. (1998). Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology*, 80(1):1–27.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, 275(5306):1593 – 1599.

Schutte, I., Slagter, H. A., Collins, A. G. E., Frank, M. J., and Kenemans, J. L. (2017). Stimulus discriminability may bias value-based probabilistic learning. *PLoS ONE*, 12(5):e0176205.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(3):461–464.

Seibt, B., Häfner, M., and Deutsch, R. (2007). Prepared to eat: how immediate affective and motivational responses to food cues are influenced by food deprivation. *European Journal of Social Psychology*, 37(2):359–379.

Seth, A. K. and Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371:2016007.

Shabat-Simon, M., Shuster, A., Sela, T., and Levy, D. J. (2018). Objective Physiological Measurements but Not Subjective Reports Moderate the Effect of Hunger on Choice Behavior. *Frontiers in Psychology*, 9:750.

Shahar, N., Moran, R., Hauser, T. U., Kievit, R. A., McNamee, D., Moutoussis, M., Nspn, C., and Dolan, R. J. (2019). Credit assignment to state-independent task representations and its relationship with model-based decision making. *Proceedings of the National Academy of Sciences of the United States of America*, 116(32):15871–15876.

Sharp, M. E., Foerde, K., Daw, N. D., and Shohamy, D. (2016). Dopamine selectively remediates 'model-based' reward learning: A computational approach. *Brain*, 139(2):355–364.

Shen, W., Flajolet, M., Greengard, P., and Surmeier, D. J. (2008). Dichotomous Dopaminergic Control of Striatal Synaptic Plasticity. *Science (New York, N.Y.)*, 321(5890):848 – 851.

Sheng, Z., Santiago, A. M., Thomas, M. P., and Routh, V. H. (2014). Metabolic regulation of lateral hypothalamic glucose-inhibited orexin neurons may influence midbrain reward neurocircuitry. *Molecular and Cellular Neuroscience*, 62:30–41.

Shin, J. H., Kim, D., and Jung, M. W. (2018). Differential coding of reward and movement information in the dorsomedial striatal direct and indirect pathways. *Nature Communications*, 9(1):1–14.

Shiner, T., Seymour, B., Wunderlich, K., Hill, C., Bhatia, K. P., Dayan, P., and Dolan, R. J. (2012). Dopamine and performance in a reinforcement learning task: evidence from Parkinson's disease. *Brain*, 135(6):1871–1883.

Siep, N., Roefs, A., Roebroeck, A., Havermans, R., Bonte, M. L., and Jansen, A. (2009). Hunger is the best spice: An fMRI study of the effects of attention, hunger and calorie content on food reward processing in the amygdala and orbitofrontal cortex. *Behavioural Brain Research*, 198(1):149–158.

Simon, J. J., Wetzel, A., Sinno, M. H., Skunde, M., Bendszus, M., Preissl, H., Enck, P., Herzog, W., and Friederich, H.-C. (2017). Integration of homeostatic signaling and food reward processing in the human brain Joe. *JCI Insight*, 2(15):e92970.

Simon, S. A., De Araujo, I. E., Gutierrez, R., and Nicolelis, M. A. (2006). The neural mechanisms of gustation: A distributed processing code. *Nature Reviews Neuroscience*, 7(11):890–901.

Skinner, B. (1959). *Cumulative record*. Appleton-Century-Crofts, Inc, Cambrigde.

Skrynka, J. and Vincent, B. T. (2019). Hunger increases delay discounting of food and non-food rewards. *Psychonomic Bulletin & Review*, 29:1729–1737.

Smith, Y., Bevan, M. D., Shink, E., and Bolam, J. P. (1998). Microcircuitry of the direct and indirect pathways of the basal ganglia. *Neuroscience*, 86(2):353–387.

Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D., and Dolan, R. J. (2013). Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. *Neuron*, 80(4):914–919.

Snaith, R. P., Hamilton, M., Morley, S., Humayan, A., Hargreaves, D., and Trigwell, P. (1995). A scale for the assessment of hedonic tone. The Snaith-Hamilton Pleasure Scale. *British Journal of Psychiatry*, 167(JULY):99–103.

Sockeel, P., Dujardin, K., Devos, D., Denève, C., Destée, A., and Defebvre, L. (2006). The Lille apathy rating scale (LARS), a new instrument for detecting and quantifying apathy: Validation in Parkinson's disease. *Journal of Neurology, Neurosurgery and Psychiatry*, 77(5):579–584.

Spence, C., Okajima, K., Cheok, A. D., Petit, O., and Michel, C. (2016). Eating with our eyes: From visual hunger to digital satiation. *Brain and Cognition*, 110:53–63.

St. Onge, J. R., Abhari, H., and Floresco, S. B. (2011). Dissociable contributions by prefrontal D1 and D2 receptors to risk-based decision making. *Journal of Neuroscience*, 31(23):8625–8633.

St Onge, J. R. and Floresco, S. B. (2009). Dopaminergic Modulation of Risk-Based Decision Making. *Neuropsychopharmacology*, 34:681–697.

Stamatakis, A. M., Van Swieten, M., Basiri, M. L., Blair, G. A., Kantak, P., and Stuber, G. D. (2016). Lateral hypothalamic area glutamatergic neurons and their projections to the lateral habenula regulate feeding and reward. *Journal of Neuroscience*, 36(2):302–311.

Stauffer, W. R. R., Lak, A., and Schultz, W. (2014). Dopamine reward prediction error responses reflect marginal utility. *Current Biology*, 24(21):2491–2500.

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., and Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience*, 16(7):966–973.

Steinert, R. E., Feinle-Bisset, C., Asarian, L., Horowitz, M., Beglinger, C., and Geary, N. (2017). Ghrelin, CCK, GLP-1, and PYY(3-36): Secretory controls and physiological roles in eating and glycemia in health, obesity, and after RYGB. *Physiological Reviews*, 97(1):411–463.

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A. E., Paliwal, S., Gard, T., Tittgemeyer, M., Fleming, S. M., Haker, H., Seth, A. K., and Petzschner, F. H. (2016). Allostatic Self-efficacy: A Metacognitive Theory of Dyshomeostasis-Induced Fatigue and Depression. *Frontiers in Human Neuroscience*, 10:550.

Stephens, D. (1981). The logic of risk-sensitive foraging preferences. *Animal Behaviour*, 29(2):628–629.

Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology and Behavior*, 106(1):5–15.

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology*, 62(6):1041–1062.

Stice, E., Spoor, S., Bohon, C., and Small, D. M. (2008). Relation between obesity and blunted striatal response to food is moderated by TaqIA A1 allele. *Science (New York, N.Y.)*, 322(5900):449–52.

Stigler, G. I. J. (1950). The Development of Utility Theory. *Source: Journal of Political Economy*, 58(4):307–327.

Stopper, C., Tse, M., Montes, D., Wiedman, C., and Floresco, S. (2014). Overriding Phasic Dopamine Signals Redirects Action Selection during Risk/Reward Decision Making. *Neuron*, 84(1):177–189.

Stouffer, M. A., Woods, C. A., Patel, J. C., Lee, C. R., Witkovsky, P., Bao, L., Machold, R. P., Jones, K. T., De Vaca, S. C., Reith, M. E., Carr, K. D., and Rice, M. E. (2015). Insulin enhances striatal dopamine release by activating cholinergic interneurons and thereby signals reward. *Nature Communications*, 6(1):1–12.

Stuber, G. D., Evans, S. B., Higgins, M. S., Pu, Y., and Figlewicz, D. P. (2002). Food restriction modulates amphetamine-conditioned place preference and nucleus accumbens dopamine release in the rat. *Synapse*, 46(2):83–90.

Stuber, G. D. and Wise, R. A. (2016). Lateral hypothalamic circuits for feeding and reward. *Nature Neuroscience*, 19(2):198–205.

Sugrue, L. P., Corrado, G. S., and Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science*, 304(5678):1782–1787.

Surmeier, D. J., Ding, J., Day, M., Wang, Z., and Shen, W. (2007). D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. *Trends in Neurosciences*, 30(5):228–235.

Sutton, R. S. (1988). Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, 3:9 – 44.

Sutton, R. S. and Barto, A. G. (2017). Reinforcement Learning: An Introduction. *book*, 2:360.

Symmonds, M., Emmanuel, J. J., Drew, M. E., Batterham, R. L., and Dolan, R. J. (2010). Metabolic state alters economic decision making under risk in humans. *PloS ONE*, 5(6):e11090.

Tai, L. H., Lee, A. M., Benavidez, N., Bonci, A., and Wilbrecht, L. (2012). Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nature Neuroscience*, 15(9):1281–1289.

Tellez, L. A., Han, W., Zhang, X., Ferreira, T. L., Perez, I. O., Shammah-Lagnado, S. J., Van Den Pol, A. N., and De Araujo, I. E. (2016). Separate circuitries encode the hedonic and nutritional values of sugar. *Nature Neuroscience*, 19(3):465–470.

Thanos, P. K., Michaelides, M., Piyis, Y. K., Wang, G.-J., and Volkow, N. D. (2008). Food restriction markedly increases dopamine D2 receptor (D2R) in a rat model of obesity as assessed with in-vivo μPET imaging ([11C] raclopride) and in-vitro ([3H] spiperone) autoradiography. *Synapse*, 62(1):50–61.

Thurley, K., Senn, W., and Lüscher, H.-R. (2008). Dopamine Increases the Gain of the Input-Output Response of Rat Prefrontal Pyramidal Neurons. *Journal of Neurophysiology*, 99(6):2985–2997.

Tiedemann, L. J., Alink, A., Beck, J., Büchel, C., and Brassen, S. (2020). Valence encoding signals in the human amygdala and the willingness to eat. *The Journal of Neuroscience*, 40(27):5264–5272.

Tiedemann, L. J., Schmid, S. M., Hettel, J., Giesen, K., Francke, P., Büchel, C., and Brassen, S. (2017). Central insulin modulates food valuation via mesolimbic pathways. *Nature Communications*, 8:16052.

Toates, F. (1986). *Motivational systems*. Cambridge University Press, Cambrigde.

Toates, F. M. (1981). The Control of Ingestive Behaviour by Internal and External StimuliA Theoretical Review. *Appetite*, 2(1):35–50.

Tobler, P. N., Christopoulos, G. I., O'Doherty, J. P., Dolan, R. J., and Schultz, W. (2009). Risk-dependent reward value signal in human prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 106(17):7185–90.

Tobler, P. N., Fiorillo, C. D., and Schultz, W. (2005). Adaptive Coding of Reward Value by Dopamine Neurons. *Science*, 307(5715):1642–1645.

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., and Schultz, W. (2007). Reward Value Coding Distinct From Risk Attitude-Related Uncertainty Coding in Human Reward Systems. *Journal of Neurophysiology*, 97(2):1621–1632.

Tolman, E. C. and Gleitman, H. (1949). Studies in spatial learning: VII. Place and response learning under different degrees of motivation. *Journal of Experimental Psychology*, 39(5):653–659.

Tom, S. M., Fox, C. R., Trepel, C., and Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518.

Tomer, R., Goldstein, R. Z., Wang, G. J., Wong, C., and Volkow, N. D. (2008). Incentive motivation is associated with striatal dopamine asymmetry. *Biological Psychology*, 77(1):98–101.

Tomer, R., Slagter, H. A., Christian, B. T., Fox, A. S., King, C. R., Murali, D., Gluck, M. A., and Davidson, R. J. (2014). Love to win or hate to lose? Asymmetry of dopamine D2 receptor binding predicts sensitivity to reward versus punishment. *Journal of Cognitive Neuroscience*, 26(5):1039–1048.

Torrubia, R., Ávila, C., Moltó, J., and Caseras, X. (2001). The Sensitivity to Punishment and Sensitivity to Reward Questionnaire (SPSRQ) as a measure of Gray's anxiety and impulsivity dimensions. *Personality and Individual Differences*, 31(6):837–862.

Tsetsos, K., Chater, N., and Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences of the United States of America*, 109(24):9659–9664.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Experiments in Environmental Economics*, 1:173–178.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

van der Plasse, G., van Zessen, R., Luijendijk, M. C. M., Erkan, H., Stuber, G. D., Ramakers, G. M. J., and Adan, R. A. (2015). Modulation of cue-induced firing of ventral tegmental area dopamine neurons by leptin and ghrelin. *International Journal of Obesity*, 39(12):1742–1749.

van Swieten, M. M. H., Pandit, R., Adan, R. A. H., and van der Plasse, G. (2014). The neuroanatomical function of leptin in the hypothalamus. *Journal of Chemical Neuroanatomy*, 61:207–220.

Verhagen, J. V. (2007). The neurocognitive bases of human multimodal food perception: Consciousness. *Brain Research Reviews*, 53(2):271–286.

Verhagen, L. A., Luijendijk, M. C., Korte-Bouws, G. A., Korte, S. M., and Adan, R. A. (2009). Dopamine and serotonin release in the nucleus accumbens during starvation-induced hyperactivity. *European Neuropsychopharmacology*, 19(5):309–316.

Volkow, N. D., Wang, G. J., and Baler, R. D. (2011). Reward, dopamine and the control of food intake: Implications for obesity. *Trends in Cognitive Sciences*, 15(1):37–46.

Volkow, N. D., Wang, G. J., Telang, F., Fowler, J. S., Thanos, P. K., Logan, J., Alexoff, D., Ding, Y. S., Wong, C., Ma, Y., and Pradhan, K. (2008). Low dopamine striatal D2 receptors are associated with prefrontal metabolism in obese subjects: Possible contributing factors. *NeuroImage*, 42(4):1537–1543.

von Neumann, J. and Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton Univerity Press, Princeton.

Vong, L., Ye, C., Yang, Z., Choi, B., Chua, S., and Lowell, B. B. (2011). Leptin Action on GABAergic Neurons Prevents Obesity and Reduces Inhibitory Tone to POMC Neurons. *Neuron*, 71(1):142–154.

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., Schreiber, L. R., Gillan, C., Fineberg, N. A., Sahakian, B. J., Robbins, T. W., Harrison, N. A., Wood, J., Daw, N. D., Dayan, P., Grant, J. E., and Bullmore, E. T. (2015). Disorders of compulsivity: A common bias towards learning habits. *Molecular Psychiatry*, 20(3):345–352.

Wall, N. R., DeLaParra, M., Callaway, E. M., and Kreitzer, A. C. (2013). Differential innervation of direct- and indirect-pathway striatal projection neurons. *Neuron*, 79(2):347–360.

Walton, M. E., Kennerley, S. W., Bannerman, D. M., Phillips, P. E. M., and Rushworth, M. F. S. (2006). 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.

Wang, D., He, X., Zhao, Z., Feng, Q., Lin, R., Sun, Y., Ding, T., Xu, F., Luo, M., and Zhan, C. (2015). Whole-brain mapping of the direct inputs and axonal projections of POMC and AgRP neurons. *Frontiers in Neuroanatomy*, 9:40.

Wang, G. J., Volkow, N. D., Logan, J., Pappas, N. R., Wong, C. T., Zhu, W., Netusil, N., and Fowler, J. S. (2001). Brain dopamine and obesity. *Lancet (London, England)*, 357(9253):354–7.

Wang, G.-J., Volkow, N. D., Thanos, P. K., and Fowler, J. S. (2009). Imaging of brain dopamine pathways: implications for understanding obesity. *Journal of addiction medicine*, 3(1):8–18.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. PhD thesis, King's College London.

Watson, P., Wiers, R. W., Hommel, B., and De Wit, S. (2014). Working for food you don't desire. Cues interfere with goal-directed food-seeking. *Appetite*, 79:139–148.

Weber, E. U., Blais, A. R., and Betz, N. E. (2002). A Domain-specific Risk-attitude Scale: Measuring Risk Perceptions and Risk Behaviors. *Journal of Behavioral Decision Making*, 15(4):263–290.

Weber, E. U., Shafir, S., and Blais, A. R. (2004). Predicting Risk Sensitivity in Humans and Lower Animals: Risk as Variance or Coefficient of Variation. *Psychological Review*, 111(2):430–445.

Weismüller, B., Ghio, M., Logmin, K., Hartmann, C., Schnitzler, A., Pollok, B., Südmeyer, M., and Bellebaum, C. (2018). Effects of feedback delay on learning from positive and negative feedback in patients with Parkinson's disease off medication. *Neuropsychologia*, 117:46–54.

Wise, R. A. (2004). Dopamine, learning and motivation. *Nature Reviews Neuroscience*, 5(6):483–494.

Wright, N. D., Symmonds, M., and Dolan, R. J. (2013). Distinct encoding of risk and value in economic choice between multiple risky options. *NeuroImage*, 81:431–440.

Wunderlich, K., Range, A., and O'Doherty, J. P. (2009). Neural computations underlying action-based decision making in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, 106(40):17199–17204.

Wunderlich, K., Smittenaar, P., and Dolan, R. J. (2012). Dopamine Enhances Model-Based over Model-Free Choice Behavior. *Neuron*, 75(3):418–424.

Xu, A. J., Schwarz, N., and Wyer, R. S. (2015). Hunger promotes acquisition of nonfood objects. *Proceedings of the National Academy of Sciences of the United States of America*, 112(9):2688–2692.

Xue, G., Lu, Z., Levin, I. P., Weller, J. A., Li, X., and Bechara, A. (2009). Functional dissociations of risk and reward processing in the medial prefrontal cortex. *Cerebral Cortex*, 19(5):1019–1027.

Yacubian, J., Gläscher, J., Schroeder, K., Sommer, T., Braus, D. F., and Büchel, C. (2006). Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. *Journal of Neuroscience*, 26(37):9530–9537.

Yacubian, J., Sommer, T., Schroeder, K., Gläscher, J., Braus, D. F., and Büchel, C. (2007). Subregions of the ventral striatum show preferential coding of reward magnitude and probability. *NeuroImage*, 38(3):557–563.

Yapo, C., Nair, A. G., Clement, L., Castro, L. R., Hellgren Kotaleski, J., and Vincent, P. (2017). Detection of phasic dopamine by D1 and D2 striatal medium spiny neurons. *Journal of Physiology*, 595(24):7451–7475.

Yin, H. H. and Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 7(6):464–476.

Zalocusky, K. A., Ramakrishnan, C., Lerner, T. N., Davidson, T. J., Knutson, B., and Deisseroth, K. (2016). Nucleus accumbens D2R cells signal prior outcomes and control risky decision-making. *Nature*, 531(7596):642–646.

Zhang, J., Berridge, K. C., Tindell, A., Smith, K. S., and Aldridge, J. W. (2009). A neural computational model of incentive salience. *PLoS Computational Biology*, 5(7):e1000437.

Zigman, J. M., Jones, J. E., Lee, C. E., Saper, C. B., and Elmquist, J. K. (2006). Expression of ghrelin receptor mRNA in the rat and the mouse brain. *The Journal of Comparative Neurology*, 494(3):528–548.