LINKING REINFORCEMENT LEARNING AND COGNITIVE MAPS TO UNDERSTAND HOW THE BRAIN REPRESENTS ABSTRACT TASKS

by

Alon Boaz Baram

A dissertation submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy



Wellcome Centre for Integrative Neuroimaging

Oxford Centre for Functional MRI of the Brain

University of Oxford

January 2020

DECLARATION

I confirm that the work presented in this thesis is my own. Where the work is the product of equally-contributed joint work or where information has been derived from other sources, I confirm that this has been clearly indicated.

Oxford, 13.01.2020

Alon Baram

ABSTRACT

The terms "reinforcement learning" (RL) and "cognitive maps" are both commonly rooted in early behaviourist psychology. However, in neuroscience, these two subfields have largely progressed in parallel. In this thesis, I use ideas and techniques from one subfield to address open questions in the other.

While the theory of RL has dramatically advanced our understanding of the brain's learning algorithms, we still don't know how RL tasks are represented. Here, I take inspiration from the well-studied representations of the cognitive map in spatial tasks to investigate how abstract non-spatial RL tasks are represented and how task knowledge might be generalised to novel situations. I present converging results suggesting that the same areas that encode the structure of spatial tasks also encode the structure of abstract RL tasks. In addition, I use ideas from spatial cognitive maps to suggest novel interpretations of heavily-studied RL neural signals. Further, taking inspiration from the discovery of inference mechanisms over the structure of spatial tasks, I suggest a study, for which I already have a task and a model, that could shed light on similar structural inference mechanisms in RL tasks.

While navigation and planning in physical space have been thoroughly studied, it is not clear how animals can navigate through cognitive maps with arbitrary topology. Here I investigate spatial cognitive maps through the lens of RL formalism, and use ideas from RL to suggest a flexible and efficient planning algorithm that can be used in both spatial environments and environments with arbitrary topology.

This thesis demonstrates the progress that can be made by bringing these two subfields together, and hopefully brings us one step closer to understanding the mechanisms underlying the human capacity for flexible behaviour.

ACKNOWLEDGEMENTS

Throughout the time of my DPhil, I was acutely conscious of how lucky I am to have the life that I have in Oxford, and to do what I do. This is entirely due to the amazing people around me, who have been with me on this ride, and held my hand when I needed them to.

First is Tim Behrens, who, despite much effort to conceal and deny it, is the most caring, sensitive and supportive supervisor out there. The knowledge that he would always be there when I actually needed him was priceless, as was the opportunity to be inspired by his astounding creativity and to learn from his approach to science. He also has good ideas sometimes, which can prove useful as his student.

The other person who probably had the most impact on this thesis is Tim Muller, my first point of contact for almost anything, who is by far the best person in the world to ramble about science with. This was a joint journey, and I couldn't have asked for a better partner.

Many other people have made these years a happy time. Thanks to Jacob and Anna for sharing a vision of what life, work and work-life relations should look like, in and out of FMRIB. Thanks to James for sharing the writing time with the right combination of support and cynicism. Thanks to the many true friends I've made in the lab, and who made me feel at home: Philipp, Helen, Mona, Cam, Thomas, Emma, Veronika, David, Shirley, Avital, Sebas, Saad, Erie and everyone else in the lab throughout the years who have made me feel so lucky to be where I am.

Big thank you to the people in the Vogels lab who were my other academic family in the first couple of years: Bill, Everton, Yayoi, Jake, Bayar, Rui, Chatanya, Friedemann and everyone else in the lab. Thanks in particular to Tim Vogels who has huge shares in the fact I even started my DPhil. I missed out on him as a supervisor, but I hope to retain him as a friend.

Oxford has been very kind with the friends it gave me, and I'll name only a few here. Thanks to Yossi for always being up for having fun. To Nomi for being the loveliest person and providing healthy dose of Israeli cynicism about it all. To my Merton gang, especially from my Masters year: Myrto, Sam, Suzanne, Marion, Marie and Samanda, who made my move to Oxford a fun and less intimidating adventure. To Ido for bringing to Oxford the main thing I was missing here, and to Ma'ayan for making me part of their journey. There

are many other people that made this place my home, and I hope they know how much I appreciate them.

A special thanks to Rafi Malach, who is the main reason why I ended up trying my luck in neuroscience, and whose enthusiastic words were always there in the background.

This DPhil would not have happened without the never-ending, loving encouragement and support of my family – Gilad, Gerry, dad and mum. I feel incredibly grateful and lucky to have them as my unshakeable base. Special thanks to mum, my number one fan, who has actually read this entire thesis and made sure I use the present perfect correctly.

To Stace, for being her, and sharing that with me. The many happy moments with her are now my definition of home.

Last, but not least, to the guys at the JR sandwich shop (and especially the blone/orange haired woman with the glasses), who helped me to find my true creative calling and facilitated the creation of my undying DPhil legacy (**Figure 0.1**). So long, and thanks for all the fish.



Figure 0.1 The "Alon Special". Possibly the best sandwich in the world.

CONTENTS

Part 1: Introduction 16

1 Reinforcement learning: a history of dichotomy 18

1.1 Formation of habits: Conditioning 18

1.1.1 Modelling associations: the Rescorla-Wagner model 20

1.1.2 Second-order conditioning 21

1.2 Cognitive maps and goal-directed behaviour 22

1.2.1 Origins – Gestalt psychology 22

1.2.2 Tolman and Purposive Behaviourism 22

1.2.3 Cognitive maps in non-space: sensory preconditioning and reinforcer devaluation 24

1.3 Models of Reinforcement Learning 26

1.3.1 Introduction to RL as an engineering subfield 26

1.3.2 The Markov Decision Process (MDP) and the Bellman equation 28

1.3.3 Model-Free temporal-difference (TD) models 30

1.3.4 Model-based (MB) algorithms 36

1.3.5 Combination and arbitration between MF and MB 38

1.3.6 Model-sensitive algorithms: the grey zone between MF and MB. 41

2 RL representations in the brain 46

2.1 Dopamine neurons activity reflect TD prediction error 48

2.1.1 Prediction error coding in fMRI 50

2.2 Evidence for multiple dissociated learning systems, and action representations in dorsal striatum 52

2.2.1 Evidence for parallel model-free and model-sensitive learning systems 53

2.3 Value representations in the ventral striatum 56

2.3.1 The role of the amygdala in calculating expected value representations 57

2.4 Model-sensitive representations in the hippocampus 58

2.5 Value representations in vmPFC/mOFC (BA 14) 59

2.6 State and learning set representations in OFC and vlPFC....64

2.6.1 Central OFC (BA 11 & 13) encodes subjective value primarily in model-based tasks, perhaps mirroring human vmPFC BOLD signals....65

2.6.2 OFC and vlPFC during learning....67

2.6.3 "The mysterious OFC"....69

2.6.4 State representations, task structure and learning sets....71

3 Cognitive maps of relational spaces in the hippocampal formation, ventral PFC and the default mode network..... 75

3.1 A cognitive map of physical space....76

3.1.1 The hippocampal zoo....76

3.1.2 The Neural underpinnings of spatial generalisation: remapping experiments....78

3.1.3 The hippocampal spatial code: more than an allocentric map....81

3.2 Non-spatial cognitive maps in the medial temporal lobe and ventral PFC....84

3.2.1 Relational memory and the hippocampal formation....84

3.2.2 Place and grid-like non-spatial representations....86

3.2.3 Other features of hippocampal non-spatial tasks maps....90

3.3 Hippocampal formation and vPFC: similarities, differences and interactions....92

3.3.1 Schemas....93

Part 2: Experimental Work..... 98

4 Physical space as a state-space with grid-like eigenspectrum, and benefits for intuitive planning 99

4.1 Introduction.... 100

4.2 The structure of 2D physical space.... 102

4.2.1 Casting space under a state space framework.... 102

4.2.2 The covariance structure of 2D physical space.... 103

4.2.3 Grid cells as an eigen-representation of the structure of 2D space.... 105

4.3 Intuitive planning: global navigation through cognitive maps using grid-like codes 108

4.3.1 Global distances from local learning.... 109

4.3.2 Global navigation from local computation.... 113

4.3.3 Navigation using intuitive planning 114

4.4 Discussion 115

5 The entorhinal cortex abstracts and generalises the structure of a reinforcement learning task 119

5.1 Introduction 119

5.2 Methods 121

5.2.1 Subjects 121

5.2.2 Training and task 121

5.2.3 Behaviour modelling 124

5.2.4 fMRI data acquisition and preprocessing 126

5.2.5 RSA analyses 126

5.3 Results 129

5.3.1 Subject behaviour 130

5.3.2 Entorhinal representations generalise across tasks with the same structure but not those with different structures. 133

5.4 Discussion 136

5.5 Supplementary information 137

6 Structure-sensitive reinforcement learning neural signals in the ventral prefrontal cortex 139

6.1 Introduction 139

6.2 Methods 141

6.2.1 Univariate analyses 142

6.2.2 RSA analyses 143

6.3 Results 145

6.3.1 The brain represents and uses a behaviourally useful latent state-space 145

6.3.2 vmPFC and ventral striatum represent the relational structure in learning signals 150

6.4 Discussion 152

6.5 Appendix 155

6.5.1 A note regarding multiple comparisons corrections in this chapter 155

7 Inferring dynamic relational structures under uncertainty 157

7.1 Introduction 158

7.2 Methods 159

7.2.1 Task 159

7.2.2 Formal description of Bayesian Learner 161

7.2.3 Selector model 168

7.3 Results 168

7.3.1 Bayesian learner 168

7.3.2 Subjects' behaviour 172

7.3.3 Supplementary figures 174

7.4 Discussion 175

8 General discussion and future directions 176

8.1 A unifying computational framework: hierarchical embedding spaces for task states 176

8.2 Open questions and future work 178

Appendix – NeuroRisk 181

LIST OF FIGURES

Figure 1.1 Example puzzle boxes. 19

Figure 1.2 Tolman and his inspiration from Gestalt psychology. 24

Figure 1.3 The reinforcer devaluation paradigm. 26

Figure 1.4. Goal-directed and habitual strategies in the devaluation task. 39

Figure 1.5 The 2-step task. 41

Figure 2.1 The reward network of the brain 47

Figure 2.2. Dopamine neurons reflect the TD prediction error in a blocking task .. 49

Figure 2.3 Parallel and dissociated learning systems for S-R ("response") and map-like ("place") learning. 54

Figure 2.4 suggested mapping of the actor-critic model to structures of the basal ganglia 57

Figure 2.5 The ventromedial PFC. 60

Figure 2.6 Value comparison signals in the vmPFC 61

Figure 2.7 Patients with vmPFC lesions are selectively impaired in a valuation task where the relationship between elements is what matters. 64

Figure 2.8 The orbitofrontal cortex. 65

Figure 2.9 Credit assignment deficits in vlPFC-lesioned animals. 69

Figure 2.10 Differences and similarities between primate and rodent OFC. 71

Figure 2.11 OFC codes the current (latent) state in an abstract task. 74

Figure 3.1 The hippocampal zoo. 77

Figure 3.2 Generalisable neural codes: remapping experiments. 80

Figure 3.3 Behaviourally-dependent changes in place fields. 82

Figure 3.4. Memory spaces and their connection to physical space. 86

Figure 3.5 Generalised spatial representations. 89

Figure 3.6 Anatomical connectivity of the DMN. 94

Figure 3.7 DMN represents stories with the same schema similarly. 96

Figure 4.1 Casting physical space as a state space. 103

Figure 4.2 Constructing the place cell covariance from local observations. 105

Figure 4.3 Grid firing fields forming a square lattice emerge from a square environment. 107

Figure 4.4 Grid firing fields forming a triangular lattice emerge from a hexagonal environment. 108

Figure 4.5 Construction of local and global knowledge from locally-learnt eigenvectors. 111

Figure 4.6 Communicability measures in arbitrary topologies. 112

Figure 4.7 Approximately linear relationship between intuitive distance and true distance. 113

Figure 4.8 Navigation in arbitrary topology worlds. 115

Figure 4.9 Failure to navigate the shortest path. 115

Figure 5.1 Task description. 130

Figure 5.2 Subjects behaviour. 133

Figure 5.3 The relational structure of the task is represented in the entorhinal cortex. 135

Figure 5.4 Control contrasts for the EC relational structure effect 138

Figure 6.1 Reward network signals reflect RL algorithmic variables operating in a state-space with the correct latent structure. 147

Figure 6.2 OFC represents a latent variable, respecting the task structure. 148

Figure 6.3 Possible latent variable repetition enhancement effects in the OFC and vlPFC. 150

Figure 6.4 Prediction error signals in vmPFC and ventral striatum depend on the current relational structure of the task. 151

Figure 6.5 Subject-by-subject correlation between the prediction error x structure interaction effect size at the vmPFC peak of the that effect and the within-block, latent variable effect size at the mOFC (BA14) peak of the effect 152

Figure 6.6 Exploratory results of the prediction error x structure interaction effect.. 156

Figure 7.1 Trial structure.. 161

Figure 7.2 Graphical description of the correlated probability tracking problem. . 169

Figure 7.3 Bayesian models with a (discretised) continuous outcome probability space 172

Figure 7.4 Parameter estimates for models with the true, 2-state parameter space for the outcome probabilities. 174

Figure 0.1 Map of the human cortex, i.e. the board in NeuroRisk. 182

Figure 0.2 Neurorisk in progress. Below: 182

ABBREVIATIONS

ACC	Anterior cingulate cortex
ANN	Artificial neural network
BA	Brodmann Area
BOLD	Blood oxygen level dependent
CS	Conditioned stimulus
CSF	Cerebral spinal fluid
DA	Dopamine
DLS	Dorsolateral striatum
DMN	Default mode network
DMS	Dorsomedial striatum
DS	Dorsal striatum
EC	Entorhinal Cortex
EPI	Echo planar imaging
fMRI	Functional magnetic resonance imaging
FSL	FMRIB's software library
FWE	Family-wise error
GLM	General linear model
HPC	Hippocampus
HRF	Hemodynamic response function
MB	Model-based
MEG	Magnetoencephalography
MF	Model-Free
MNI	Montreal Neurological Institute
mPFC	Medial prefrontal cortex
mOFC	Medial orbitofrontal cortex
MDP	Markov Decision Process
MRI	Magnetic resonance imaging
MS	Model-sensitive
MTL	Medial temporal lobe
MVPA	Multivariate pattern analysis
NA	Nucleus accumbens
OFC	Orbitofrontal cortex
PCA	Principal component analysis

PCC	Posterior cingulate cortex
PE	Prediction error
PFC	Prefrontal cortex
PHCg	Parahippocampal gyrus
PIT	Pavlovian to instrumental
RL	Reinforcement Learning
RNN	Recurrent neural network
ROI	Region of interest
RSA	Representational similarity analysis
RSC	Retrosplenial cortex
RT	Response time
RPE	Reward prediction error
SNpc	Substantia nigra pars compacta
SPE	Structure prediction error
SPM	Statistical parametric mapping
S-S	Stimulus-Stimulus
S-R	Stimulus-Response
SR	Successor Representation
TE	Echo time
TR	Repetition time
TPJ	Temporal-Parietal Junction
TD	Temporal difference
US	Unconditioned stimulus
vPFC	Ventral Prefrontal cortex
vmPFC	Ventromedial prefrontal cortex (BA14)
vStr	Ventral Striatum
VTA	Ventral tegmental area
VTE	Vicarious trial and error

Part 1: Introduction

"I always get to where I'm going by walking away from where I have been"

Winnie-the-Pooh, a bear of very little brain

*Our brains have evolved to optimise action selection in order to increase our chances to reproduce and survive. Making successful choices, even those that seem trivial, is a hard problem: the world is complex and changing, and there are usually countless alternative courses of action. How does the brain perform this monumental feat? The computational theory of Reinforcement Learning (*RL (Sutton and Barto, 1998)*) has proved a highly successful framework for explaining a wealth of learning and decision-making behavioural results, as well as elucidating the brain's algorithms for learning.*

Critically, RL *algorithms assume knowledge about what information is important for the task at hand. They depend on appropriate, low-dimension representations of the high-dimensional sensory input that the animal is constantly bombarded with* (Behrens et al., 2018b; Niv, 2019; Walton et al., 2010; Wang et al., 2018; Wilson et al., 2014). *Pens and pencils should be represented similarly when learning to write, but differently when learning to start fires. By focusing on different features (dimensions) of the world (here nibbiness vs woodiness), the respective novices (humans, animals or models) can learn faster in each domain. Discovering what is a good representation of a task can be a hard and data-consuming problem* (Gershman and Niv, 2010; Wang et al., 2018).

However, humans and animals can solve seemingly complex tasks with little or no direct experience. For example, Harry Harlow (Harlow, 1949) *asked primates to repeatedly choose between a changing pair of two stimuli, only one of which was rewarded. His subjects quickly learned what the crucial feature of the stimuli was: whether or not they were rewarded before. This feature is not tied to the particular sensory properties of the stimuli. It is relational, enabling Harlow's subjects to use their learned task representation to generalise their knowledge to a choice between two novel*

stimuli (Harlow, 1949)[1]. *Such impressive behaviour suggests good task representations must exist in the brain. However, we currently know very little about their nature.*

One related, yet distinct area of neuroscience has described its own task representations in exquisite detail – the cognitive map of physical space (O'Keefe and Nadel, 1978; Tolman, 1948). *These representations provide examples for how some of the representational problems we want solved in RL are solved in space, for example generalisation. In this thesis, I will take inspiration from cognitive maps to try and get a better understanding of RL task representations. I will also use the RL framework to investigate navigation through cognitive maps.*

In Chapter 1 of this introduction, I review the rich psychological literature suggesting (at least) two behavioural strategies (as well intermediate strategies) are used by animals and humans, namely habitual (model-free) and goal-directed (model-based) strategies. I then present the standard RL computational algorithms and discuss their strengths and weaknesses. In Chapter 2, I present evidence for the dependence of RL processes on specific brain structures and discuss identified neural representations of these processes. Specifically, I highlight the gap in our knowledge about the nature of representations underlying complex state-spaces and model-sensitive/based decision-making. In Chapter 3, I discuss the highly-related notion of cognitive maps and its neural correlates, where representations have been described in exquisite detail. Specifically, I will discuss how relational structures might be represented in the hippocampal formation, ventral PFC and other areas in the default mode network. Recently, these areas, on which cognitive maps rely, have also been suggested as underlying complex state-spaces and model-sensitive strategies (Garvert et al., 2017; Miller et al., 2017; Schuck et al., 2016; Stachenfeld et al., 2017; Vikbladh et al., 2019; Wilson et al., 2014).

[1] Note that for the results Harlow's original experiment there are other, simpler explanations than the acquiring of complex learning sets: a simple model-free agent with a learning rate of 1 (Chapter 1.3.3.1) will implement a win-stay, lose-shift strategy required by Harlow's task. However, it is possible to view the learning of the right learning rate as the learning of the learning set in this case.

1 REINFORCEMENT LEARNING: A HISTORY OF DICHOTOMY

"All students agree as to the facts. They disagree, however, on theory and explanation" (Tolman, 1948).

Choices can be made based on more than one strategy. This seemingly intuitive idea was at the heart of a fierce debate in the psychology community of the early 20th century. Specifically, the debate focused on a dichotomy between two classes of control: habitual and goal-directed (Dolan and Dayan, 2013). *In the first two parts of this chapter I explore this dichotomy from the point of view of its rich psychological literature. In the last part of the chapter I review important corresponding topics from the parallel computational field of reinforcement learning.*

1.1 Formation of habits: Conditioning

Edward Thorndike (Thorndike, 1911, 1905, 1898) is perhaps the first psychologist who sought to identify laws of behaviour based on rigorous experimental procedures and systematic description to the study of learning, rather than the earlier introspective trends in psychology – a programme that would later become known as "behaviourism". In a typical experiment, Thorndike placed a hungry cat in a puzzle box, and placed food outside the box. Each box had a door that could open by triggering a specific mechanism (e.g. press a lever – pull on a string – push a bar). Through trial and error, the cat stumbled upon the right sequence of actions that opened the door and obtained the food reward. The next time the cat was put in the box, it escaped substantially more quickly. Thorndike considered his work as an investigation of associative structure between stimuli (puzzle box) and responses (set of actions that open the door), which emerged due to reinforcement (escape). This view was

summarised in his *Law of Effect*: "Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction are more firmly connected with the situation … ; those which are accompanied or closely followed by discomfort … have their connections with the situation weakened" (Thorndike, 1911). In other words, Thorndike explained the behaviour of his cats as resulting from reinforcement[1] of the association between stimulus (S - puzzle box) and response (R - sequence of actions leading to escape). He claimed his work "has rejected reason, comparison or inference…" as explanations for the results (Thorndike, 1898). "The fundamental phenomena which I find present in animal consciousness… emphasise the fact that our (human) mental life has grown up as mediation between stimulus and reaction" (Thorndike, 1898).

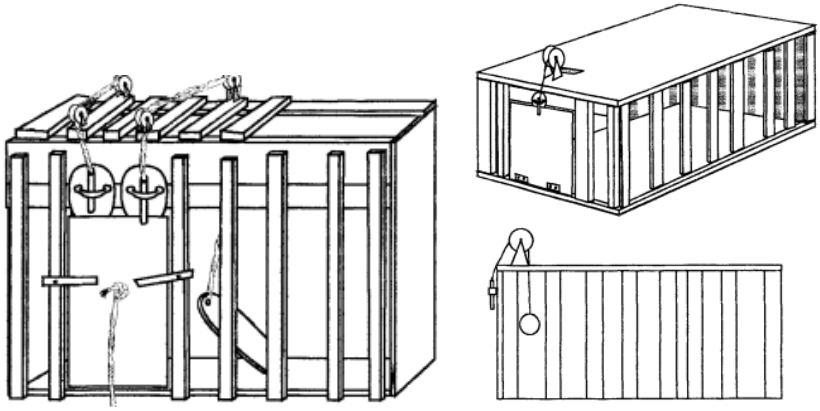


Figure 1.1 Example puzzle boxes. Adapted from (Thorndike, 1898)

Around the same time in St. Petersburg, Russia, another founding father of modern psychology was investigating the digestion system of dogs. Ivan Pavlov first noted that the presence of food (unconditioned stimulus or US) made his dog salivate more (unconditioned response). He then paired an originally neutral stimulus (ringing of a bell –the conditioned stimulus or CS) with the US many times and measured the response (amount of salivation – conditioned response). While before the pairing (termed "*conditioning*") there was no increase in response when the bell rang, after the pairing there was – even in the absence of any food. The difference between the Pavlovian (or classical) conditioning and Thorndike's

[1] Thorndike did not actually use the term "reinforcement" (which was used by Pavlov) until 1933 (Thorndike, 1933)

instrumental conditioning is that in the former the delivery of reward following the stimulus is not contingent on the animal's behaviour. In other words, in Pavlovian and instrumental conditioning the S-S or S-R association are being reinforced, respectively. A generation of celebrated psychologists like James Watson and Clarke Hull unified and formalised these two types of conditioning under the framework of S-R Behaviourism (Hull, 1943). Some behaviourists, like B.F. Skinner (founder of "Radical Behaviourism") highlighted the critical differences between Pavlovian and instrumental (termed by Skinner "operant") conditioning.

1.1.1 Modelling associations: the Rescorla-Wagner model

In this thesis, I will use several computational models of behaviour to help characterise behaviour and explain neural signals. Hence, in this introduction I will review the most influential of these models. I start by describing the Rescorla-Wagner model, originally suggested as an explanation for the blocking effect.

In the late 1960s the view that pairing on its own was enough to establish an association was challenged, for example by *blocking* experiments (Kamin, 1969). In such an experiment, a CS *A* (e.g. light) is first paired with a US (e.g. food) many times, forming a Pavlovian association and resulting in conditioned responses to *A*. Then, *A* is once again paired with the US many times, but now this is done in the presence of another CS, *X* (e.g. tone). Though *X* was paired with the US, there was no conditioned response to it when presented on its own: it was "blocked" by the associative power of *A*.

While earlier models treated each CS separately, assuming they acquired an independent value, the *Rescorla-Wagner model* explains the changes in the CS-US associations in terms of the effectiveness of the US: How surprising is the US, or equivalently how well is the US predicted by the CS. Because *A* fully predicts the US, experiencing the *US* after *AX* is not surprising, and hence the *X-US* association does not get strengthened.

Formally, the Rescorla-Wagner model states that:

$$\Delta V_X^{n+1} = \alpha_X \beta (\lambda - V_{total}^n)$$
$$V_X^{n+1} = V_X^n + \Delta V_X^{n+1}$$

Equation 1.1

Where ΔV_X^{n+1} is the change in the associative strength V of CS X as a result of pairing with the US at trial $n+1$, $0 < \alpha_X, \beta < 1$, are constants, representing the associability

(usually meaning salience or intensity) of the CS X and the US, respectively, λ is the maximum associative strength that the US can support in any single situation (e.g. the amount of reward) and V_{total}^{n} is the sum of associative strengths of all CSs presented in trial $n+1$. It is evident from these equations that the magnitude and sign of the surprise (i.e. the "*prediction error*") to the presentation of the US ($\lambda - V_{total}^{n}$) determine the dynamics of the model. Though originally formulated to account for Pavlovian conditioning, the Rescorla-Wagner model can easily be expanded to include instrumental conditioning by tracking the associative strength of actions instead of stimuli. The term "associated strength" can be thought of as the *associative value* of the CS – how well does it predict reward.

The scalars α_X and β, which control the degree by which the prediction error updates the estimate of the associative value, were later combined and reinterpreted in the machine learning community as the "learning rate" of the model. Together with the more widespread interpretation of λ as the amount of reward obtained, we can rewrite the Rescorla-Wagner model (in the case of only a single CS) as:

$$V_X^{n+1} = V_X^n + \alpha(r^n - V_X^n) = (1-\alpha)V_X^n + \alpha r^n$$

Equation 1.2

Where α is the learning rate and r^n is the reward obtained on trial *n*. We will encounter extensions to this model in sections 1.3.3 and 5.2.3, demonstrating its centrality in behavioural neuroscience and reinforcement learning.

1.1.2 Second-order conditioning

The Rescorla-Wagner model had many successes, but also failures. A particularly important one, which we will encounter later in sections 1.3.3, is the case of second-order conditioning. Here, CS *A* is first paired with a reinforcer (e.g. food) to asymptote. Then *A* is presented for several trials with another CS, *X*, this time without reinforcement. According to the Rescorla-Wagner model, $\lambda - V_{total}^{n}$ (the prediction error) will be negative, as $V_{total}^{n} = V_A^n > 0$. This means that the associative value of both *A* and *X* will decrease, and in the case of *X* will become negative – i.e. *X* now predicts "negative food". While such cases can indeed be observed (termed "conditioned inhibition" (Rescorla, 1969)), in the majority of cases the exact opposite phenomena occurs – the animal responds to the second CS *X* in the same way that it responded to *A* (which was originally paired with reinforcement). *X* inherits some

of *A*'s association with reward, perhaps by *A* becoming a reinforcer on its own accord – "second-order" conditioning.

1.2 Cognitive maps and goal-directed behaviour

1.2.1 Origins – Gestalt psychology

While bottom-up Behaviourist approaches to psychology were growing stronger in America, a different approach was being developed in Germany – the Kantian, top-down Gestalt psychology, founded by Max Wertheimer, Kurt Koffka and Wolfgang Köhler. Köhler studied chimpanzee behaviour when presented with complex novel problems. For example, he hung a banana from the ceiling of a room, out of reach, and placed several boxes in the room. The apes stacked the boxes one on top of the other to make use as makeshift ladder and reached the banana (**Figure 1.2**a). This behaviour could not be explained by trial and error or conditioning. Rather, the apes seemed to be able to mentally *simulate* possible behaviours they had never experienced, in order to reach their goal. They used prior knowledge about *relationships* between objects (e.g. "ontopness") to create something they never encountered before. Köhler claimed the apes had an insight, which they proceeded to carry out in a way that was "unwaveringly purposeful" (Köhler, 1917).

1.2.2 Tolman and Purposive Behaviourism

Not all behaviourists completely disagreed with the principals of Gestalt psychology. Edward Tolman, who's "Purposive Behaviourism" (Tolman, 1932) would become one of the leading approaches to psychology in the 1930s, visited Giessen in 1912 as a student, where Koffka was teaching, and even served as a subject in some early Gestalt experiments. Over the next three decades Tolman kept in continual contact with Koffka – a connection which dramatically influenced his work (Sokal, 1994).

In a remarkable set of experiments with a "behaviourist setting", summarised in the seminal paper "Cognitive Maps in Rats and Men" (Tolman, 1948), Tolman and his colleagues at Berkeley showed that rats can make flexible inferences and learn rich details of mazes in the absence of any reward – behaviour that could not be explained by simple stimulus-response associations. For example, in "latent learning" experiments, after unrewarded exposure to mazes, rats would take (never rewarded before) shortcuts to reach rewards

(Blodgett, 1929; Tolman and Honzik, 1930). In "spatial orientation" experiments, the configuration of a maze changed to reveal novel direct routes leading to a reward that stayed in an old spatial location. The rats were able to take the shortest route to the reward, even though it was never experienced (let alone reinforced) before (Tolman et al., 1946) (**Figure 1.2**b-c). Note, however, that there was a prominent shortcoming in the design of this experiment: in both training and testing periods, a light signalled the goal location, meaning it is possible that the rats used this local cue (rather than an elaborated cognitive map) to choose the "shortcut" (Olton, 1979). Indeed, a human virtual reality replication attempt of this experiment provided evidence for this dependence of subjects on this local cue in their choices (Wilson and Wilson, 2018).

Another type of behaviour highlighted by Tolman as evidence for cognitive map-like processing was "Vicarious Trial and Error" (VTEs) (Muenzinger, 1938; Tolman, 1939), a name given to the hesitating, looking back-and-forth behaviour which rats often display at a choice-point before choosing one way or the other. Tolman interpreted VTEing as evidence that "in the critical stages [of learning]… the animal's activity is not just one of responding passively to discrete stimuli, but rather one of the active selecting and comparing of stimuli" (Tolman, 1948)". In other words, VTEing is (indirect) evidence that the rats were actively *searching* through their cognitive map, and mentally *simulating* the possible outcomes of the different choices. While it was clear to Tolman that this was more sophisticated and complex process than suggested by the S-R school, computational tools to understand just *how* complex is this simulation were only available a decade later (Bellman, 1957). We explore these tools further in section 1.3.

In some sense, Tolman reinterpreted a standard behaviourist experimental setting in Gestalt terms. He agreed with the S-R advocates that rats running in the maze are exposed to stimuli, with the final result of this exposure being the responses. Indeed, he was a behaviourist. But he disagreed with the Watson's behaviourism. Tolman's Purposive Behaviourism (Tolman, 1932) highlighted the importance of the cognitive map as a mediator between the S and the R. The rats did not just respond "helplessly" to a succession of external and internal stimuli. "Intervening brain processes are more complicated… and more autonomous than do the stimulus-response psychologists [claim]" (Tolman, 1948).

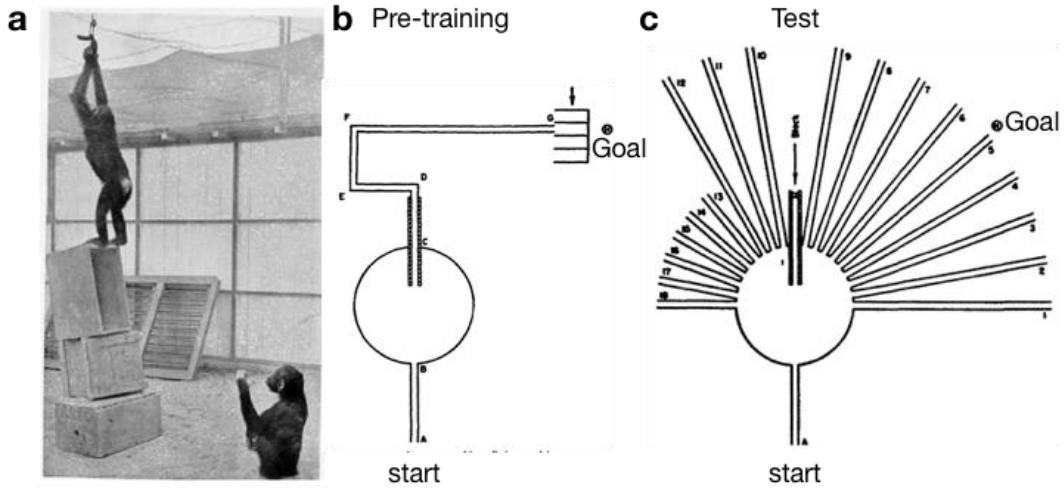


Figure 1.2 Tolman and his inspiration from Gestalt psychology. a. Monkeys showing "insight" regarding possible uses for boxes in novel situations. From (Köhler, 1917). **b.** and **c.** "spatial orientation experiments in rats: **b.** The maze rats were trained in. **c.** The maze rats were tested in. In the majority of cases rats took the arm leading directly to the goal in the first trial, even though they had never experienced this arm before. Adapted from (Tolman et al., 1946).

1.2.3 Cognitive maps in non-space: sensory preconditioning and reinforcer devaluation

Tolman did not see the cognitive map as only spatial, though all of his experiments were spatial. In principal, his theory also extends to explain the behaviour of Köhler's monkeys as using the abstract "map" of objects (e.g. boxes), physical relations (e.g. ontopness) and actions (stacking boxes to create an ontopness relationship between them). Here I will highlight two non-spatial behavioural paradigms where S-R and cognitive maps-based behavioural control produce dissociable predictions.

In a *sensory preconditioning* (Brogden, 1939) experiment two neutral stimuli (e.g. a light and a tone) are paired with each other by being presented together (Brogden, 1939; Rescorla and Cunningham, 1978) or in immediate (Cole et al., 1995; Hoffeld et al., 1958) or delayed (Ward-Robinson and Hall, 1998) succession. One of the stimuli (e.g. light) is then paired with a reinforcer to asymptote in a classical conditioning procedure. The response to the other stimuli (e.g. tone) is then tested in extinction (with no delivery of any further reward). The S-R theories predict there will not be any response to the tone, as there was never a reinforcer to strengthen the tone-response association. Converging results indicate otherwise: animals also responded to the stimulus (tone) that was never directly paired with reinforcement. This suggests that even in the absence of reinforcement, a stimulus-stimulus (S-S) association was learned – analogous to Tolman's "latent learning" spatial experiments.

The *reinforcer devaluation* paradigm (Adams and Dickinson, 1981; Balleine and Dickinson, 1998; Colwill and Rescorla, 1985; Dickinson and Mulatero, 1989) targeted another important weakness of law-of-effect type behaviours: insensitivity to the identity of reinforcer, and its current appeal to the animal. In a typical experiment (Colwill and Rescorla, 1985) action (or stimulus) A (e.g. left lever press) is paired with reinforcer X (e.g. sucrose liquid) while action B (right lever press) is paired with reinforcer Y (food pellet). Baseline choice preferences between actions A and B are measured. A is then devalued outside the context of the experiment, e.g. by pairing it with illness or by feeding it to satiety. The animals are then tested in extinction (choice between A and B, without any reinforcement). According to S-R theories, the S-R association is "stamped in" by the reinforcer (e.g. food), but the reinforcer itself is not encoded in the associative structure. Animals will "*habitually*" repeat the same response when presented with the stimulus (testing box), even if the reinforcer is not desirable any more. In contrast, a cognitive map-like approach assumes the animal can mentally simulate the consequences of each action, enabling it to make a choice based on its current *goals* (Dickinson and Balleine, 1994). The majority of studies found reliable devaluation effects, suggesting that animals act in a goal-directed manner. However, in many cases (e.g. (Colwill and Rescorla, 1985)) there is still a residual response consistent with an S-R association (choosing the action that will lead to the devalued reward). This is especially true in cases where the animals received extensive training (Adams, 1982). These results suggest both the habitual and the goal-directed systems are being used, depending on the circumstances. We explore this topic in greater detail in section 1.3.5.

As a related side note, it is important to observe that this dependence of the observed effect on (sometimes minimal) differences in the behavioural protocol of tasks is not uncommon, and can be both useful and confusing. For example, in two of the tasks discussed so far – sensory preconditioning and second-order conditioning, opposite results can be observed depending on the exact protocol. In sensory preconditioning, the amount of exposures to pairing of the neutral stimuli increases preconditioning (response to A after $\underline{B}$ has been paired with reward) initially, but this effect reverses with repeated exposers – more exposers to parings of the neutral stimuli will decrease preconditioning (Hoffeld et al., 1960). In second-order conditioning, pairing of CS1 with the US followed or interspersed with pairing of CS2 with CS1 results in "excitation" of CS2 – the animal is more likely to respond to CS2 than a control. However, if there are too many CS1-CS2 pairings, the effect is reversed and CS2 becomes "inhibitory" - animals are less likely to respond to it than to a control (Yin

et al., 1994). Both of the examples above are useful to demonstrate that care and attention to details is needed when making profound theoretical claims from behavioural data: for example, while sensory preconditioning is many times used as an example of cognitive map-like S-S learning that is similar to S-R learning without reward, the results above suggest that there is indeed a difference between S-S learning and standard conditioning.

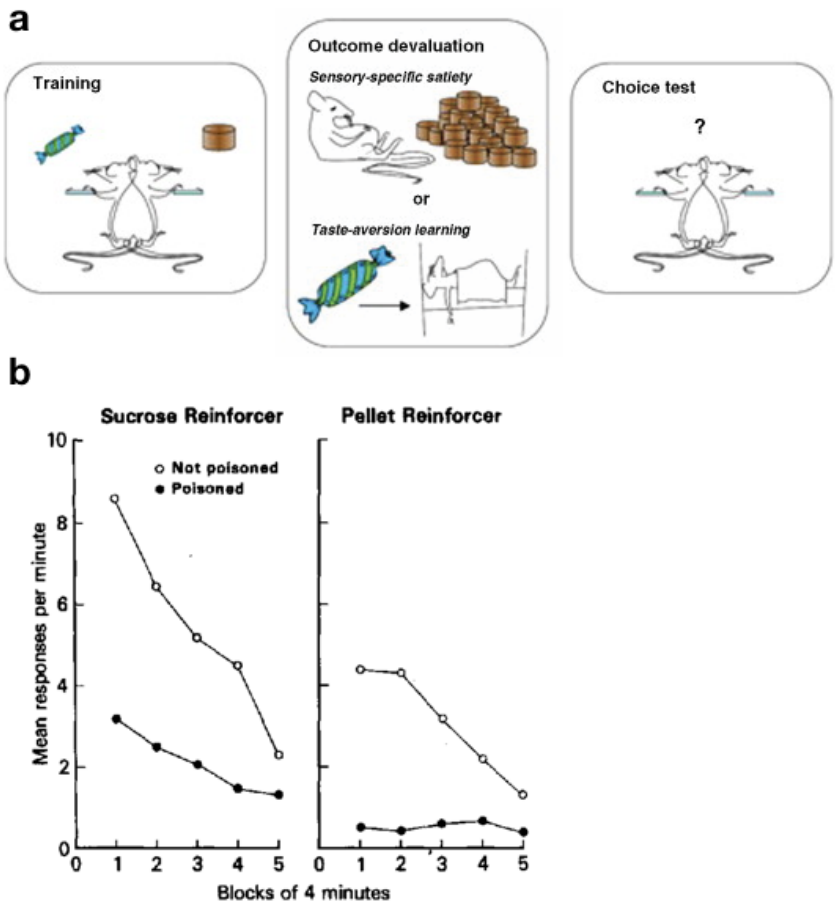


Figure 1.3 The reinforcer devaluation paradigm. a. Schematic of a typical experiment. Adapted from (Ostlund and Balleine, 2008). **b.** Data showing the devaluation effect from (Colwill and Rescorla, 1985).

1.3 **Models of Reinforcement Learning**

1.3.1 Introduction to RL as an engineering subfield

In the field of artificial intelligence, reinforcement learning (Sutton and Barto, 1998) agents are systems that learn by trial-and-error, and are "hedonistic" – they are trying to increase the chance of receiving rewards and avoiding punishments through the actions they

can choose to take. There is no "informed teacher" telling the system if their actions are correct or not. The system gets (sometimes delayed) "evaluation signals" from the environment regarding how good or bad the sequence of actions it took was – but it never knows if those were the best or worst actions to be taken. RL systems are faced with the *explore-exploit dilemma*: no matter how good your current actions are, you'll never know if there's a better action out there. These features make RL a unique branch of machine learning, distinct from supervised and unsupervised learning.

The origins of RL partly lie in the animal learning literature I have reviewed above ("reinforcement" was a term first used by Pavlov), but also partly lie in some of the earliest engineering work in artificial intelligence: a branch of applied mathematics named "optimal control", and its solution through "dynamic programming" (Bellman, 1957). Briefly, optimal control deals with designing controllers of a dynamical system such that an objective function is optimised. For example, in the case of the Uber app, the dynamical system is composed of all available drivers, ride requests and their current locations, the objective function to be minimised is the [time customers wait] + [company's expenses like fuel cost] (or more likely a weighted sum emphasising the latter), and the controller is the assignment of rides to drivers. Another example of an optimal control problem is that of a mouse trying to reach a food pellet in a spatial maze. The controller in this case is the policy (decision-making rule) of mouse. Primarily due to the separation of disciplines and their different goals (understanding animal behaviour on the one hand, building machines that perform optimally under certain constraints on the other), the connection between animal learning and optimal control research was not acknowledged until the late 1970s.

One of the earliest attempts at building a computational model of trial-and-error RL was made by Marvin Minsky, co-founder of MIT's Media (AI) Lab, in his PhD thesis (Minsky, 1954). He was also the first to formulate the "*credit assignment problem*": How should an agent distribute the credit for a reward between the many decisions that led to that reward (Minsky, 1961)? All of the RL methods I will discuss in this chapter can be seen as attempts to solve this problem. Minsky was the first to realise that second-order conditioning experiments, where animals directly face the credit assignment problem, might be useful for the purposes of artificial intelligence.

The RL problem can be divided into two, highly related problems: the prediction problem and the control problem. The *prediction problem* usually concerns predicting the value of each of the options the agent can choose. These values can change, based on feedback from the

environment. This is the same problem studied in Pavlovian conditioning experiments. The *control problem* concerns the choices the agent makes and is usually highly tied to the prediction problem: intuitively, most models will usually choose options with high predicted values.

In the next section, I first present the formal framework for the vast majority (if not all) of RL problems – the Markov Decision Process (MDP) and discuss the Bellman equation – the key computation all RL models are trying to solve or approximate. I then present examples of the two important classes of RL models: model-free (MF) and model-based (MB), which can be seen as the engineering equivalent of the S-R vs cognitive maps debate in psychology. First are the temporal difference (TD) models, a family of model-free algorithms which suggest a solution to the prediction problem and were inspired by the Rescorla-Wagner model. I also briefly discuss SARSA, Q-learning and the Actor/Critic models – model-free TD agents which suggest solutions to both the prediction and the control problems. Next, I discuss model-based algorithms and dynamic programming. Finally, I will present some models and strategies that lie within the spectrum between model-free and model-based methods, termed by (Miranda et al., 2019) "model-sensitive" methods. Throughout the section I will also (incompletely) discuss the successes and failures of these models in explaining behavioural data.

Note that in this section I only present formal descriptions of models and demonstrate examples where they proved useful in explaining behavioural data. In the next chapter we will see how these models have been astoundingly useful in elucidating the brain's decision-making processes.

1.3.2 The Markov Decision Process (MDP) and the Bellman equation

An MDP is a mathematical abstraction that captures many features of the type of sequential tasks RL agents are trying to solve. It consists of a set of possible states the environment ("the world") might be in ($\mathcal{S}$), a set of actions the agent can choose, ($\mathcal{A}$) and a scalar reward function which depends on the states ($\mathcal{R}$). At any given (discrete) time point t, the environment takes on a state $s_t \in S$, and delivers a reward based on that state $r_t(s_t = s)$

(denoted $r_t(s)$ for simplicity)[1]. The agent then chooses an action $a_t \in \mathcal{A}$, and the world transitions to the next time point $t+1$. The next state of the world s_{t+1}, depends (perhaps stochastically) only on the previous state and action (s_t, a_t). This is that Markov property that gives the MDP its name. In other words, the next state is chosen according to the transition distribution $T = P(s_{t+1}|s_t, a_t)$. A state-space of a task which is not inherently Markovian (e.g. because it has some latent features) can easily be made Markovian by duplicating the observable states to accommodate the latent features (this is further explained in section 1.3.6.3). For the purposes of this thesis (and indeed almost all RL tasks), I assume that the spaces states and action are finite.

The agent's goal in each time step is to choose the action which maximises the expected future rewards. RL agents usually solve this by first solving (or approximating) the prediction problem: they compute a value function over the states. This can be written as:

$$V(s) = \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots | s_t = s]$$

Equation 1.3

This expression is also known as the expected *return* (when starting at state s) – the discounted future reward (with a discounting factor $0 < \gamma < 1$, prioritising more immediate rewards). From here on I will use the more common notation of s' , s'' etc. to denote the successor states of s. Equation 1.3 is somewhat misleading in its simplicity, as the expectation is taken over possible actions and future states. Importantly, future states depend on the agent's actions. Hence, to unpack this equation, we first need to unpack the expectations over actions. We define the *policy* $\pi(s): S \rightarrow A$ as a mapping from a given state to the actions (or distribution over actions) the agent takes in this state. Informally, the value of state s under policy π is the expected return when starting in s and following policy π. We can now rewrite Equation 1.3 using only expectations over future states (as expectations over actions are incorporated in the policy):

$$V^{\pi}(s) = r(s) + \gamma\, \mathbb{E}[r(s')]_{s'} + \gamma^2 \mathbb{E}[\mathbb{E}[r(s'')]_{s''}]_{s'} + \ \ldots \ =$$

[1] This delivery can be stochastic, i.e. according to a distribution $P(r_t|s_t)$, in which case I will denote the expectation (over rewards) of this distribution as r_t.

$$r(s) + \gamma\, \mathbb{E}[r(s') + \gamma \mathbb{E}[r(s'') + \ \dots\ ]_{s''}]_{s'}$$

Where $\mathbb{E}[]_{s'}$ denotes the expectation over all possible successor states $s' \in \mathcal{S}$ Crucially, the nested structure of this expression allows it to be written in a recursive form:

$$V^{\pi}(s) = r(s) + \gamma\, \mathbb{E}[V^{\pi}(s')]_{s'}$$

Equation 1.4

We can equivalently make the same calculation by explicitly writing down the expectations:

$$V^{\pi}(s) = r(s) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s))(r(s') + \gamma \sum_{s'' \in \mathcal{S}} P(s''|s', \pi(s'))(r(s'') + \gamma \dots))$$

$$= r(s) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s)) V^{\pi}(s')$$

Equation 1.5

This form of the value function (under the policy π) is known as the *Bellman equation* (Bellman, 1957). Solving it provides a solution to the prediction problem, at least under a fixed policy.

1.3.3 Model-Free temporal-difference (TD) models

In section 1.1.1 I discussed (with slightly different notation) the Rescorla-Wagner (RW) model where the value estimate trial $t + 1$ was calculated by using the prediction error $\delta = r_t - \hat{V}(S_t)$ (where $\hat{V}$ denotes the model's estimation of the value, which should be distinguished from the true value V) to "nudge" the current value estimate "in the right direction". This direction is defined according to the current reward, in order for the value to converge the expected reward ($\hat{V}(S_t) = \mathbb{E}[r_t]$). $\delta = 0$ in expectation only when the prediction of $\hat{V}(S_t)$ is correct: it measures to what extent the equality $\hat{V}(S_t) = \mathbb{E}[r_t]$ holds on any particular trial. We can now apply the same trick to the Bellman equation (Equation 1.4): for it to converge, the predicted values and the expected rewards should match up, i.e. $\hat{V}(S_t) = r_t + \gamma \mathbb{E}[\hat{V}(S_{t+1})]$ (again we assume $r_t = \mathbb{E}[r_t]$ for simplicity). We can measure to what extent this equality holds in any particular trial by substituting the expected value of the next state with the model's current estimation of that value for the observed next state, and measure the discrepancy (error) between the two sides if the equation:

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

Equation 1.6

This is the temporal-difference prediction error. The origins of the name are evident from the formula. In trials where there is no direct reward obtained, the prediction error is exactly the (discounted) difference between temporally successive states. The value estimation can then be updated in the same way as in the Rescorla-Wagner model (Equation 1.1):

$$\hat{V}(s_t) \leftarrow \hat{V}(s_t) + \alpha\delta$$

Equation 1.7

Changes in $\hat{V}(s_t)$ when direct reward is not obtained occur when state transitions produce changes in reward expectation, which should then propagate and update previous expectations. In these states, TD learns from (assigns credit through) a prediction error between its *own (temporally successive) predictions,* rather than an error between the actual outcome and the prediction[1] (like other machine learning MDP algorithms such as Monte Carlo methods (Sutton and Barto, 1998)).

Some TD algorithms (e.g. TD(λ) (Sutton, 1988)) use the temporal difference between state values that do not immediately succeed each other but are rather several time steps in the future (equivalently the past, see the discussion about the equivalence of the "forward" and "backward" views of TD(λ) in (Sutton and Barto, 1998)). In these algorithms, the values of all previously visited states are updated in every transition (rather than just the current state), with the amount of update depending on the current prediction error and an eligibility trace reflecting how recently each state was visited (Sutton and Barto, 1998). This allows for the propagation of value updates after receiving reward to occur faster than just one timestep at a time.

Note that the difference between the TD and RW equations is the addition of future expected rewards term $\gamma \hat{V}(S_{t+1})$. This highlights the conceptual difference between the

[1] The method of using the model's own estimation is called Bootstrapping.

models: the TD model was designed to solve *sequential tasks* (MDPs), whereas the RW model was designed to solve trial-wise conditioning problems, or more broadly to track the associative value of specific objects (like stimuli or actions). In TD, learning occurs continuously from moment to moment, and is not confined to the time-frame of trials.

However, it turns out the TD model actually does a better job in modelling behaviour in conditioning experiments than RW (and indeed was inspired by RW (Sutton and Barto, 1981)) in cases where relative timings of experimental events matter (Ludvig et al., 2012). It preserves all the successes of the RW model like the ability to predict blocking (section 1.1.1) but addresses some of the RW model failures. For example, recall that one of the shortcomings of the RW model is its failure to predict second-order conditioning (section 1.1.2). In TD, during training where stimulus A predicts a reward, $\hat{V}\left(S_{t=[time\ of\ presentation\ of\ A]}\right)$ will become high as it predicts the reward (as well as the values of all states at the timepoints between the presentation of A until the reward is presented). From here it's easy to see from Equation 1.6 how A functions as a "surrogate reward" for its preceding states, such that when subsequent training on $B \rightarrow A$ occurs, the high value propagates also to the earlier state $S_{t=[time\ of\ presentation\ of\ B]}$.

1.3.3.1 Making choices using TD

So far, I have discussed how TD methods learn to predict the values of states given a fixed policy. In this section I present a brief and incomplete account of TD-based methods to optimise this policy in order to maximise the expected return. All of these models have theoretical convergence guarantees under reasonable assumptions and with enough time, they will find the optimal policy. They rely on computation of the value functions in the spirit of the previous section, and some also include a computation of a separate policy function. Combined with powerful function approximating techniques that have become available in recent years in the form of Deep Neural Networks, these methods have proven remarkably useful in solving complex and high-dimensional problems (e.g. (Mnih et al., 2015; Silver et al., 2016). However, these algorithms have several critical shortcomings, which generally map to the shortcomings of S-R behaviourism theories. I discuss these further after presenting the algorithms.

1.3.3.1.1 SARSA

Equation 1.5 presents the Bellman equation, calculating the value of states, assuming the actions that will be taken in the state follows a fixed policy Π. The same equation can be

formulated for state-action pairs, i.e. calculating the expected return for taking action a in state s, and then following the policy Π thereafter. To differentiate the state-action value from the value of states (V), we denote them by Q:

$$Q^{\Pi}(s,a) = r(s) + \gamma \sum_{s' \in S} P(s'|s,a) Q^{\pi}(s', \Pi(s'))$$

Equation 1.8

This is the Bellman equation, stated for state-action pares rather than just states. The agent learns these values from experience in the same way as explained before: when a state transition occurs following the choice of an action, the values are updated based on a prediction error (Equation 1.7) that uses the reward obtained and the estimates of the values of the next state-action pair:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$$

The actions (in both current and next trial, which determine which values will participate in the update of the current state-action value) are chosen based on a policy that is derived from Q. For example, the agent can just take the action with the maximal Q-value. A smarter comparison method uses a softmax function that is weighted by the Q-values, such that the agent is most likely to choose the action with the highest value but it can also *explore* other options:

$$P(a|s) = \frac{exp(\beta Q(s,a))}{\sum_{a' \in \mathcal{A}(s)} exp(\beta Q(s,a'))}$$

Equation 1.9

The sequence state-action-reward-state-action, which is involved in each update of the values, give the algorithm its name (SARSA). It can be proved that the policy resulting from this strategy is equal or better than the policy Π that defines it, resulting in convergence to the optimal policy, denoted Q^*.

1.3.3.1.2 Q-learning (off-policy learning)

The update equation for the Q-learning algorithm is very similar to the SARSA algorithm, except that in the former, the "next action" used for the Q-value update does not need to be the action the agent actually takes:

$$\delta = r(s) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) argmax_a Q^{\pi}(s', \mathrm{a}') - Q^{\Pi}(s,a)$$

This trick is called "*off-policy*" learning. It enables a more accurate learning of the value function (as it uses a better estimate for future rewards – the value of the *best* action to be taken at the next state but still allows the agent the freedom to explore the other, less-good actions, in the choice it actually makes). This is a crucial feature that enables better handling of the explore-exploit dilemma. Once again, the Q-learning algorithm uses a sampling to approximate this equation, using TD methods:

$$Q^{\Pi}(s_t, a_t) \leftarrow Q^{\Pi}(s_t, a_t) + \alpha r_{t+1} + \gamma argmax_{a_{t+1}} Q^{\pi}(s_{t+1}, \mathrm{a}_{\mathrm{t+1}}) - Q^{\Pi}(s_t, a_t)$$

1.3.3.1.3 The Actor/Critic model

The actor/critic model has two parts, corresponding to the prediction and control problem. The "adaptive critic" is a TD-learner of state values, learned with the same methods presented in section 1.3.3: it calculates the TD value prediction error using Equation 1.6 and updates the state values using Equation 1.7. The second part of the actor/critic model, the "actor", separately keeps track of state-action values, though it learns these through prediction errors generated by the critic (unlike SARSA and Q-learning):

$$\omega(s_t, a_t) \leftarrow \omega(s_t, a_t) + \eta\delta$$

Equation 1.10

and then uses these state-action values to define a probabilistic choice policy, usually using a softmax function analogous to Equation 1.9. Importantly, in Equation 1.10 δ is the prediction error generated by the critic (Equation 1.6), meaning the critic "trains" the actor. The actor uses a separate learning rate η – different than α used by the critic (Equation 1.7).

1.3.3.2 TD as a successful formal description of S-R behaviour, and its consequent weaknesses as a learning algorithm

All of the TD agents presented in section 1.3.3.1 choose actions based on a value that is assigned to a state-action pair, based on reinforcement. We can identify the RL abstractions "state", "action" and (immediate) "reward" with Thorndike's "situation" (by which he meant the stimuli or sensory events the animal experiences, and that can easily be extended to internal "events"), "response" and "satisfaction" in his formulation of "the law of effect" (section 1.1). TD algorithms formalise Thorndike's original ideas for proximal, and extend them to sequential tasks by including the future expected reward as part of the "satisfaction". The essence of the computation is the same for Thorndike and the TD algorithms: a mapping

from states to responses (Daw, 2016). This comparison is also useful for a discussion of the cases where TD algorithms do not perform well, and to cases where they do not match the behaviour of animals.

Perhaps the greatest challenge in contemporary RL is the amount of experience (data) needed for learning, and the difficulty of generalisation. These are the same shortcomings of habitual, S-R behaviour. TD algorithms are highly restricted by their training data , and usually can't generalise learning to novel situations (though there are some interesting semi-exceptions, like (Wang et al., 2018), which we discuss further in section 1.3.6.2). They have to learn how to act in the new situation from experience. In contrast, human and animal behaviour is flexible. We can choose how to act by exploiting actions that have worked in the past but also based on experiences that are only loosely related; we can imagine the consequences of entirely novel choices. We can abstract important features of experiences and generalize them to new situations (Behrens et al., 2018b). As explained in section 1.2, S-R based theories like TD-learning cannot explain how Tolman's rats found never-before-experienced shortcuts, nor how Köhler's monkeys imagined a new use for boxes as building blocks for a make-shift ladder. As TD-algorithms only learn from reward and will only choose actions that previously led to reward, they could never explain how Tolman's rats knew immediately how to run to the reward in the latent learning experiments.

Another example of a situation where S-R methods fail is the reinforcement devaluation paradigm (section 1.2.3). Specifically, TD-algorithms have no representation of the reward – it is only used to learn "cached" state-action values. Hence, when the reward function changes (e.g. reward *X* is devalued outside the context of the task), the TD learner has no mechanism for updating these values except through direct experience[1].

[1] A possible (though unappealing) solution for simple, non-spatial devaluation paradigm within the TD framework is to "duplicate" the state-space and learn the task separately for the different rewards, and to include the "non-task" context where the devaluation occurs as another (directly experienced) state in the task with negative reward. In such a case, the obvious question is where does such a sophisticated state-space representation come from, and how does the animal know to represent the task in such a way. I further discuss how sophisticated state-space representations might help in learning complex tasks through simple model-free algorithms in 1.3.6.3.

1.3.4 Model-based (MB) algorithms

Two shortcomings of TD-models were highlighted by latent learning and reinforcer devaluation experiments: the former highlighted their inability to learn in the absence of reward, and the latter highlighted their inflexibility to changes in the reward function, respectively. Both of these shortcomings are addressed in a different class of RL models, called *model-based* (Daw et al., 2005; Sutton and Barto, 1998) or *model-sensitive* ((Miranda et al., 2019), see section 1.3.6) models. Historically, the invention of these algorithms as engineering tools chronologically preceded the algorithms presented so far. Model-based algorithms are so named because they learn an *internal model* of the environment (the MDP), by representing the two defining features of the MDP: the immediate reward associated with each state ($r(s)$), and the transition function $P(s_{t+1}|s_{t,}a_t)$). Both can be learnt easily: $r(s)$), with the Rescorla-Wagner algorithm, and $P(s_{t+1}|s_{t,}a_t)$ by counting the experienced transitions between states. Appropriately, models that do not store an internal model of the environment, such as TD models, are called *model-free* (MF) models. MF models usually only store a "cached" state-action value function.

The first and most straight-forward MB algorithm does not attempt to learn $r(s)$ and $P(s_{t+1}|s_{t,}a_t)$ – it assumes complete and accurate knowledge of them. This case is true in many engineering scenarios, yet the solution to the optimal control problem (finding the optimal policy) still needs to be found. This can be performed by a class of algorithms called *dynamic programming* (Bellman, 1957). If $r(s)$) and $P(s_{t+1}|s_{t,}a_t)$ are known, the value function (and consequently the policy it implies) can be directly computed by iterative expansion (into future states) of the recursion of the Bellman equation. This results in a "tree" structure of states (where the "children" nodes of state s are all the states that can be transitioned to). The value of the final state in the recursion can be directly experienced, and values of all preceding states (including accumulation of rewards along the way) can then be calculated "up the tree" through the nested sums in the top equation of Equation 1.5.

A standard example of a dynamic programming algorithm is the *value iteration* algorithm (Sutton and Barto, 1998): initialise the state-action value function to zero, and then go state by state and update the state value according to the (one-step) Bellman equation using the current value estimates of all other states:

$$Q^*(s,\mathrm{a}) \leftarrow r(s) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) argmax_{a'} Q^*(s',\mathrm{a}')$$

Equation 1.11

Once all states have been updated, start again from the first state (this time using the updated value estimates of all other states. If this is repeated for enough iterations, it is guaranteed to converge. This process is akin to "simulating" the consequences of future actions, sometimes referred to as "tree-search". MB algorithms perform this search when choosing each action. Hence, they compute their policy from a full, "correct" representation of $r(s)$ and $P(s_{t+1}|s_{t,}a_t)$, making them flexible to changes in these functions. If $r(s)$ and $P(s_{t+1}|s_{t,}a_t)$ are not given, the same computations can be performed on estimates of them learned from experience (though in such cases the choice policy might bias the estimation).

Such flexibility is what is required from a model in order to produce cognitive map-like and goal-directed behaviours, which MF learners fail to produce (section 1.3.3.2). "Latent learning" behaviour (Gläscher et al., 2010; Tolman, 1948), i.e. learning without reward, is achieved by learning and storing $P(s_{t+1}|s_{t,}a_t)$, even in the absence of reward (for example through a random walk policy or passive observation). Similarly, the ability to find never-experienced-before routes when the configuration of the MDP changes can be achieved by updating $P(s_{t+1}|s_{t,}a_t)$ to the current correct configuration (or generalising/inferring it from previous encounters with problems with a similar structure (Behrens et al., 2018b)) and simulating paths to the reward (in Tolman's "spatial orientation" experiments (section 1.2.2) favourite arms are blocked and previously blocked arms are opened, and animals can presumably update $P(s_{t+1}|s_{t,}a_t)$ to incorporate these changes through visual input – without any need to experience the transitions). Goal-directed behaviour like that observed in the reinforcer devaluation paradigm can be explained by using Equation 1.11 to calculate values with the updated reward function $r(s)$. Interestingly, the hidden assumption made here – that an animal always has access to the reward function in a devaluation paradigm because the value depends on the animal's own motivational state – does not always hold. Under certain conditions (mainly novel motivational states that the animal does not have a lot of experience with), animals need to experience the outcome in the new motivational state before they update the reward function. This process is termed incentive learning (Balleine and Dickinson, 1991).

MB algorithms calculate values on-the-fly and are thus superior to MF algorithms in terms of the accuracy of the value estimation whenever a change in the environment has occurred. However, this comes at a great computational cost. The number of computations that need

to be performed grows exponentially with the number of states that need to be simulated into the future. This is what Bellman called "*the curse of dimensionality*" (Bellman, 1957), and it renders dynamic-programming a useful algorithm only in task with a relatively simple and shallow state-space.

1.3.5 Combination and arbitration between MF and MB

MF and MB methods differ dramatically: MF models only store cached values, and need substantial sampling of the environment in order to learn them. They are therefore, like behavioural habits, inflexible to changes in the reward or transition function. MB models use their stored representation of the environment to plan and are therefore adaptive to environmental change. The computations MF models perform at every choice are inexpensive and relatively simple. MB methods perform expensive simulations of possible trajectories in the future. Assuming these are the only methods available (an assumption we challenge in section 1.3.6, though it has generally been prevalent in most research in the field in the last 15 years), it is sensible that animals would employ both strategies, and choose the best one depending on the situation. Behavioural data from revaluation experiments[1] demonstrate this concept: with moderate amounts of training, animals show the goal-directed devaluation effect (**Figure 1.4**a, left). However, with extensive training this effect disappears, i.e. animals tend to show habitual control. The devaluation effect remains even after extensive training in more complex tasks – for example if the animal needs to perform two actions in order to get two different rewards ((**Figure 1.4**a, right, (Adams, 1982)).

[1] This is a slightly different variant of the devaluation paradigm than the one shown in section 1.2.3 from (Colwill and Rescorla, 1985). Here, there isn't a choice between two different actions that lead to two different outcomes (devalued and non-devalued), as in (Colwill and Rescorla, 1985). Rather, there is only one outcome which is either devalued or not, and the choice facing the animal is whether to press the lever or to enter the food magazine to look for the reward.

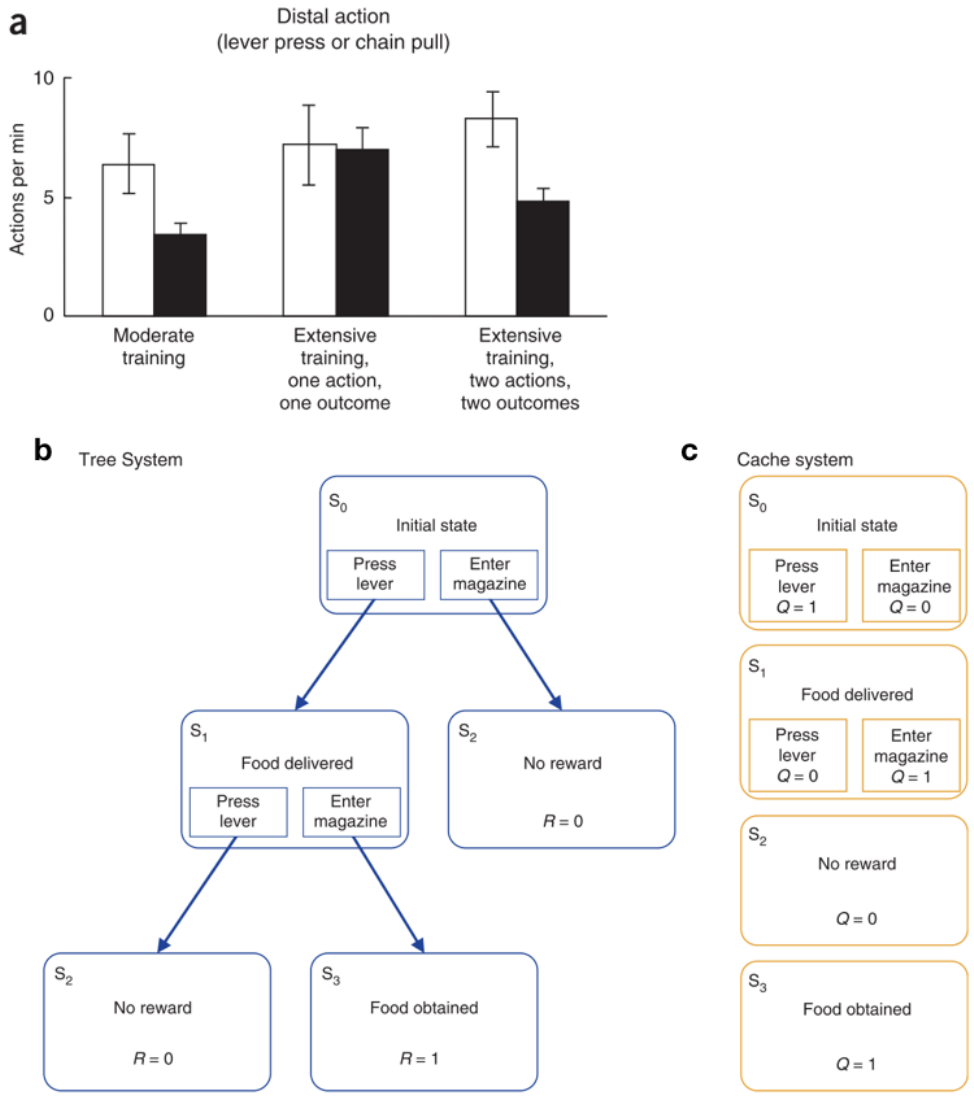


Figure 1.4. goal-directed and habitual strategies in the devaluation task. a. Results showing dependency of shift to habitual behaviour on the amount of training. Adapted from (Daw et al., 2005), who in turn adapted from (Adams, 1982). **b. and c.** Devaluation task representations using a MB model (tree search, **b.**) and a MF model (Q-learning, **c.**). Adapted from (Daw et al., 2005).

(Daw et al., 2005) showed this behaviour can be obtained as a result of uncertainty-based arbitration process between a MB model (tree search, i.e. dynamic programming, Figure 1.4b) and a MF model (Q-learning, Figure 1.4c).

The same authors (give or take) later extended this framework and presented a task where it is possible to obtain a behavioural signature indicating whether MF, MB or (crucially) MF-MB *combination* strategy is being used by subjects (Daw et al., 2011). The task later received the name "*The 2-Step Task*" given its structure (Figure 1.5a) and has since become the standard task for differentiating MF and MB behaviour. At the first step the subject chooses between two actions, termed action A and action B. Subsequently the subject reaches one of

two second-step states, termed state a and state b. Action A leads to state a 70% of trials, and action B leads to state b 70%; Importantly, on a randomly selected 30% of trials, a rare transition occurs, such that action A leads to state b and action B to state a. In each second-step state two further actions are available. The subject chooses one of these actions and either receives or not a reward. The reward probabilities for each of the four second-step actions (two in each second-step state), drift over time (Figure 1.5b).

The task was designed so that a MF and MB algorithms will show different behaviour. While a MF model should only care about the reception of reward in each trial (regardless of the statistical structure of the task between steps 1 and 2), a MB model will learn the transition probabilities between steps 1 and 2 and use them to simulate future outcomes to inform choices in step 1. This difference can be seen in Figure 1.5c and Figure 1.5d: the probability of choosing the same choice as the last trial ("stay probability") for a MF learner depends only on whether a reward was given in the last trial (Figure 1.5c), while for a MB learner this probability depends also on whether the transition observed in the last trial was common or rare: rewards increase stay probability following common transitions and decrease stay probability following rare transitions. Daw and colleagues found that the behaviour of subjects corresponded to a hybrid model (Gläscher et al., 2010), combining the two models by a weighted sum of their outputs.

This task has been used extensively since, for example to show that in humans, the balance between the MF and MB systems can be tipped towards the former with manipulations like cognitively demanding tasks (A Ross Otto et al., 2013) or with stress (A. R. Otto et al., 2013).

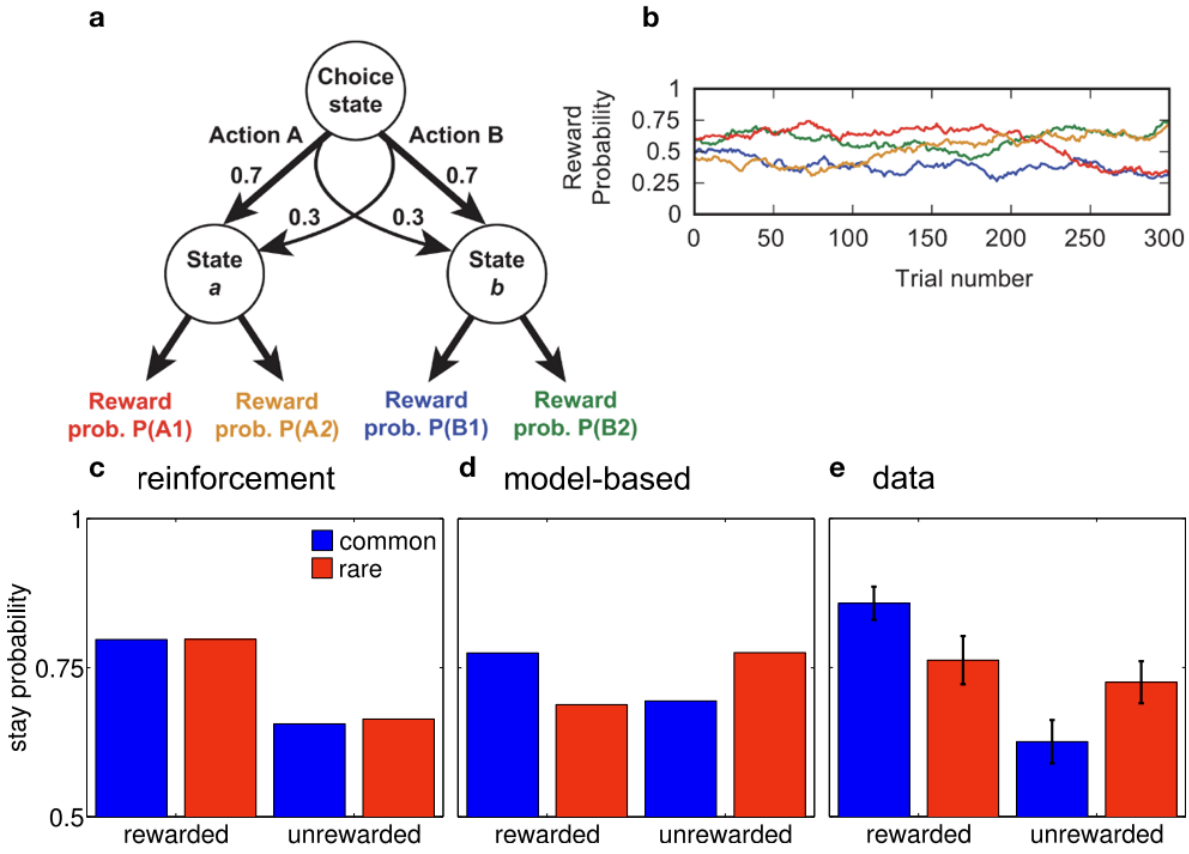


Figure 1.5 The 2-step task. a. Task structure. **b.** Drifting reward probabilities associated with each of the four final states. a. and b. adapted from (Akam et al., 2015). **c-e.** Probability repeating a choice in the first step, as a function of the previous trial (weather it was rewarded and whether the transition between the step 1 and 2 was common or rare. **c.** Prediction of a MF model. **d.** Prediction of a MB model. **e.** Data from humans. c-e. adapted from (Daw et al., 2011).

1.3.6 Model-sensitive algorithms: the grey zone between MF and MB.

In this section we will explore some methods that lie in the spectrum of statistical efficiency and computational simplicity between MF and MB extremes, termed by Peter Dayan (Miranda et al., 2019) "*Model-sensitive*" (MS) methods. MS models include any models that exploit information associated with the task dynamics, in contrast with MF models that simply repeat actions that were good in the past. In this sense, MB methods can be seen as a specific example of MS methods. Here we use the definition for MB given by (Miranda et al., 2019) – methods that use the knowledge of the task dynamics to explicitly plan, for example by simulating (usually through tree search). Non-MB MS methods (which I'll refer to as just "MS methods") use their knowledge about the task structure in ways that are less computationally expensive and are not subject to "the curse of dimensionality". Though

historically RL has mostly focused on MF and MB methods, MS methods have recently received a lot of attention in psychology and neuroscience.

Several MS methods arise from interactions between MF and MB methods (Daw and Dayan, 2014). These include integrating MF values into depth-limited MB planning (Keramati et al., 2016), MF control over MB calculations (Cushman and Morris, 2015; Dayan, 2012; Lieder et al., 2018) and MB training of MF (Sutton, 1991). Here, I focus on two other MS methods that are of particular interest to this thesis: learning (through TD) and caching a predictive representation of the transition structure of the task (Dayan, 1993), and the construction of sophisticated model-dependent representations of task states that enable MF methods to learn efficiently (Akam et al., 2015). I also mention a recent Recurrent Neural Network (RNN) model which, despite learning through MF mechanisms, has had some success in representing and generalising the task structure in its recurrent dynamics (Wang et al., 2018).

1.3.6.1 The successor representation

The successor representation (SR, not to be confused with S-R (stimulus-response)) was proposed by Peter Dayan in 1993 (Dayan, 1993) as a method to represent the task states in a way that preserves the intuitive notion that successive states should have a similar representation. If such representations can be learned and cached (i.e. so that they won't need to be computed from scratch at every trial, like in MB methods), this should enable fast generalisations between similar states. The SR caches long-term predictions about the states the agent is expected to visit in the future (unlike MB methods that just store the one-step transitions), a process sometimes referred to as temporal abstraction. Namely, for each starting state the SR caches how often the agent expects to visit each of its future successor states. The SR is then combined with the reward function to calculate state values, following the critical observation that the value function can be decomposed into an inner product of these two functions:

$$V(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t) \,|s_0 = s] \;=\; \sum_{s' \in \mathcal{S}} M(s, s') R(s')$$

Equation 1.12

where the SR M is defined as

$$M(s, s') = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \mathbb{I}(\, s_t = s') |\, s_0 = s] = \sum_{t=0}^{\infty} \gamma^t T^t$$

Equation 1.13

Where $\mathbb{I}$ is the indicator function (1 when true, 0 otherwise) and T is the transition matrix. This means M simply counts how many times (in expectation) the agent will visit state s' when starting in state s, and discounts this count according to how further in the future this visit will occur. Importantly, the SR can be learnt incrementally using temporal difference updates, by replacing the reward prediction error with a "successor" prediction error (Dayan, 1993).

From Equation 1.12 we can see that the SR retains important benefits of both MF and MB algorithms: Like MB methods, it is flexible to changes in the reward function (as at decision time it linearly combines its cached SR representation with the *current* reward function). Like MF methods, it does not perform a tree search simulation every time a choice needs to be made, and is hence much less laborious than MB. SR-based action evaluation has similar computational complexity to MF, while retaining some of the flexibility of MB.

However, the SR does not retain all the flexibility that MB can benefit from. It is not flexible to changes in the transition structure, like the ones that occurred when maze arms where blocked or opened in Tolman's spatial orientation experiments. (Momennejad et al., 2017) harnessed this dissociation in flexibilities between MF, MB and SR-based methods to ask which strategies are used by humans in a task where either the reward function, the transition function, or both changed. Results indicated that indeed there was differential sensitivity of subjects to reward vs transition structure changes as predicted by use of the SR, though subjects did show some flexibility to the transition structure as predicted by MB.

It is also worth noting that as the SR learns the transition function from experience, it is policy-dependent: the action the agent takes will influence the transition it experiences, which in turn will influence the SR. This is unlike "pure" MB methods like value iteration, which assume full and correct knowledge of the transition structure and calculate state-action values for all actions that can be taken in a given state.

1.3.6.2 Meta-RL

As mentioned in section 1.3.3.2, the success of harnessing the advances of deep Artificial Neural Networks (ANNs) to RL problems has been, though impressive, limited in its ability to generalise to situations that the models were not explicitly trained on. In general, deep RL models employ MF learning techniques: they use the reward prediction error (RPE) as the objective function for the ANN to minimise, and learn complex mappings between inputs

and actions. (Wang et al., 2018) used such a learning procedure to train a Recurrent Neural Network (RNN) to perform tasks which are commonly used in the RL psychology literature. Interestingly, a seemingly insignificant trick they used has resulted in an ability to generalise training on a domain of tasks to a novel instance of that domain: the network receives the reward obtained and action taken in the previous timestep as an input, together with the task-specific inputs and the recurrent inputs. The recurrent network weights are trained using the RPE. Crucially, a separate and independent RL algorithm emerges in the recurrent activation dynamics of the RNN (even when there is no synaptic, RPE-driven learning happening – the weights are fixed). This algorithm can solve RL tasks in a manner that shows evidence of knowledge about the task structure (the things common to the tasks it was trained on) – even if the specific task it solves is novel. Hence, the authors gave the network the attractive name "meta-RL". The emergent algorithm shows the characteristics of a MB model, for example in the 2-step task. This suggests that MB control could arise, at least in part, by an MF-like, RPE-based training procedure.

1.3.6.3 The importance of a useful state representation

The algorithms discussed above, whether MF, MB or MS, are all designed to learn how to make good choices in a given MDP – i.e. given a set of states, a set of actions and a reward function. A hidden assumption of these models is that the state-space is given to the agent. Moreover, RL assumes that the agent is magically endowed with a *useful* state-space - that describes the task "correctly", out of the many state-spaces that might describe the task. This is a very non-trivial assumption: the animal is bombarded with a high-dimensional constant stream of external and internal inputs and needs to somehow arrive at a low-dimensional task space representation on which it can use its smart RL learning algorithms for learning. A good state-space representation means that learning which actions to take in each state *could* actually help the agent to better perform at the task (Niv, 2019; Wilson et al., 2014). Note that the usefulness of a state-state-space depends also on the learning/solving algorithm - different state-spaces of the same task might be differentially suited for different algorithms.

A state can be defined as a summary of the features of the external and internal environment that are relevant for future behaviour in a particular task (Langdon et al., 2019). For example, not all behaviours are stimulus bound. States should incorporate autonomously arising internal information relevant for the task, and subjective features like context and memory. With a good state-space representation of the task, complex behaviours can be

achieved by simple algorithms. For example, (Akam et al., 2015) challenged the wide-spread interpretation of the 2-step task results from (Daw et al., 2011) (**Figure 1.5**), by showing that when operating on a sophisticated state representation[1], a model-free agent shows behaviour that is qualitatively very similar to the "hallmark" behaviour of a model-based agents.

How the state-space is learned, sometimes referred to as "structure learning" (Acuña and Schrater, 2010; Tenenbaum et al., 2011), is an open question. In chapters 5 and 6, I offer an incomplete answer to this question: when possible, the brain reuses (generalises) state-spaces it has already learned for previously encountered problems with a similar structure. While this solution does not explain how the initial learning occurred (Gershman et al., 2015; Gershman and Niv, 2010; Niv, 2019; Tomov et al., 2018), it does suggest at least a partial solution to the state-space representation/structure learning question in the vast majority of problems humans and animals face.

[1] The states of the second step were defined on the basis of latent task relevant information: it separates the step 2 task states *a* and *b* into 2 states each, based on whether reward was obtained in the last trial. The MF agent learned action values which produced the strategy of choosing action A following rewards in state a, action B following rewards in state b, action B following no reward in state a and action A following no reward in state b. This corresponds to the classic MB behaviour - strong stay probability transition-reward interaction.

2 RL REPRESENTATIONS IN THE BRAIN

In chapter 1 I reviewed the rich psychological history of the decision-making field, discussed key ideas in the machine learning field of reinforcement learning, and gave examples where this theory provided parsimonious explanations to behavioural experimental phenomena. RL has had an even greater influence on the field of decision-making neuroscience, and our understanding of the brain's mechanisms for learning. In this chapter, I review these key contributions, focusing on neural representations that have been successfully interpreted as RL algorithmic variables.

The vast majority of this progress centres on model-free learning algorithms in the brain. Hence, I start by reviewing the striking mappings between elements of these algorithms (particularly the actor-critic model (Joel et al., 2002)*) and the basal ganglia. These include the famous correspondence between midbrain dopaminergic signals and the temporal-difference prediction error, as well as representations of state values in the ventral striatum (vStr). Together, these representations constitute the "critic". I also discuss recent challenges to this canonical model, including observations that the dopaminergic prediction error carries more information than just a pure reward prediction error. In the actor-critic framework, the dorsolateral striatum (DLS) is the "actor" (also trained by dopaminergic prediction errors), harbouring a representation of the policy over actions and of the choice history. I then use the particularly striking double dissociation between the lateral and medial dorsal striatum, coding for habitual and goal-directed policies respectively, to demonstrate that the brain, at least to some extent, follows the dichotomy between MF and MB learning described in the previous chapter.*

Next, I briefly discuss the role of the hippocampus (thoroughly discussed in the next chapter) in model-based behaviour, before turning to the prefrontal cortex, starting with the ubiquitous value representations found in the ventromedial PFC (vmPFC) which seem to be tuned to structural aspects of tasks. Finally, I focus on a question that has been highlighted only in recent years, and which has implications for all types of RL algorithms and neural representations – the representation of the

task state space. Specifically, I will focus the role of the ventral PFC (and particularly OFC) in this representation and discuss how this role can be reconciled with the area's value representations.

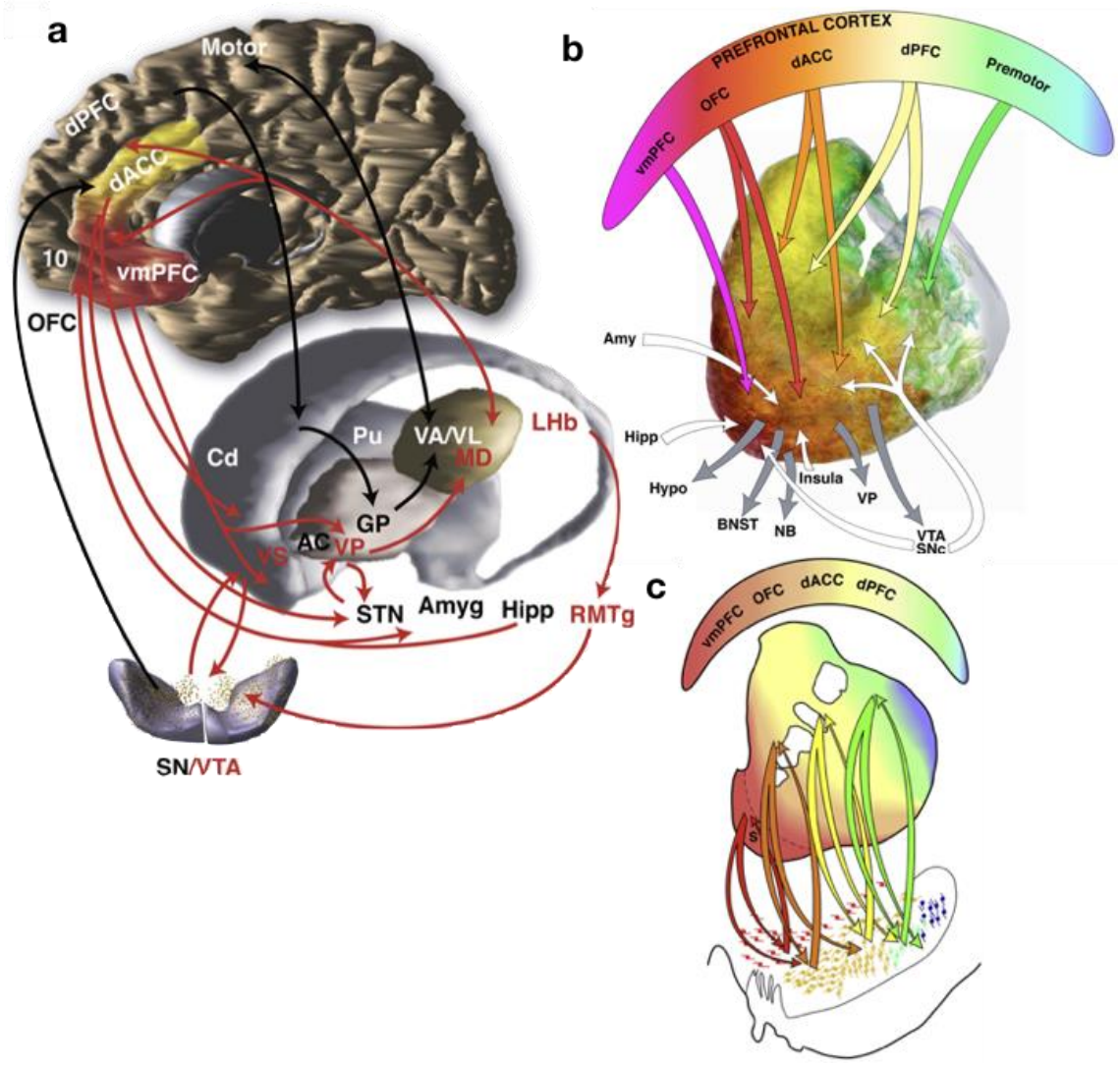


Figure 2.1 The reward network of the brain. a. The place of the reward network within the basal ganglia. **b.** Illustration of the organisation of striatal connections. **c.** Illustration of the organisation of the connections between the striatum and the substantia nigra. Structures that are discussed in this thesis: Amyg: Amygdala; dACC: dorsal anterior cingulate cortex. dPFC: dorsal prefrontal cortex; Hipp: Hippocampus; OFC: Orbitofrontal cortex; Pu: putamen; SN: Substantia nigra, pars compacta; vmPFC: ventromedial PFC; VS: ventral striatum; VTA: ventral tegmental area. Structures not discussed in this thesis: Cd: caudate nuclei; GP: globus pallidus; LHb: lateral habenula; RMTg: rostromedial tegmental nucleus; STN: subthalamic nucleus; VA/VL/MD: ventral anterior/ventrolateral/medial-dorsal nuclei of the thalamus; VP: ventral pallidum. From (Haber and Behrens, 2014)

2.1 **Dopamine neurons activity reflects TD prediction error**

Dopamine (DA) is a neuromodulator released from midbrain neurons in the substantia nigra pars compacta (SNpc) and the ventral tegmental area (VTA), which, through extensive and elaborated axon arborisation, innervate large parts of the striatum and of the frontal cortex. DA can be released either following action potentials of the dopaminergic neurons (phasic release) or as background, tonic release. Here I focus on the former, though it is suggested that tonic DA regulates the response to phasic DA by setting the baseline for DA receptor stimulation. Throughout the 1990s, a series of reports regarding the responses of DA neurons in conditioning experiments and rewards in general culminated in an influential synthesis suggesting the experiments can be modelled using the TD model, and that the phasic activity of DA neurons tracks one of the critical quantities of the model – the reward prediction error (RPE) (Montague et al., 1996; Schultz et al., 1997).

Take, for example, a blocking paradigm (section 1.1.1), where a stimulus A is trained to predict reward, while stimulus B is trained to not predict reward. Stimulus X is presented together with A while stimulus Y is presented together with B, with both compounds predicting reward. Learning on stimulus X is "blocked", as the reward is fully predicted by A, but learning on stimulus B (the control) is not blocked. (Figure 2.2a). These task and behavioural results, as explained in section 1.3.3, are successfully modelled by the TD model.

Dopamine responses during outcome (reward/no-reward) following the presentation of X and Y on their own reflect prediction errors from the beliefs that X does not predict reward while Y does (Figure 2.2b). To understand the responses during cue times, it is important to note that stimuli presentations occur with random time intervals between them, so that when a stimulus was presented, the animal did not expect any reward. Hence, if reward-predicting stimuli (A, AX, BY, Y) are being presented, DA neurons will increase their activity (positive prediction error Figure 2.2c top left, middle left, middle right, bottom right, respectively), while presentation of stimuli that do not predict reward (X, B) results in no change of activity (Figure 2.2c bottom left, top right, respectively).

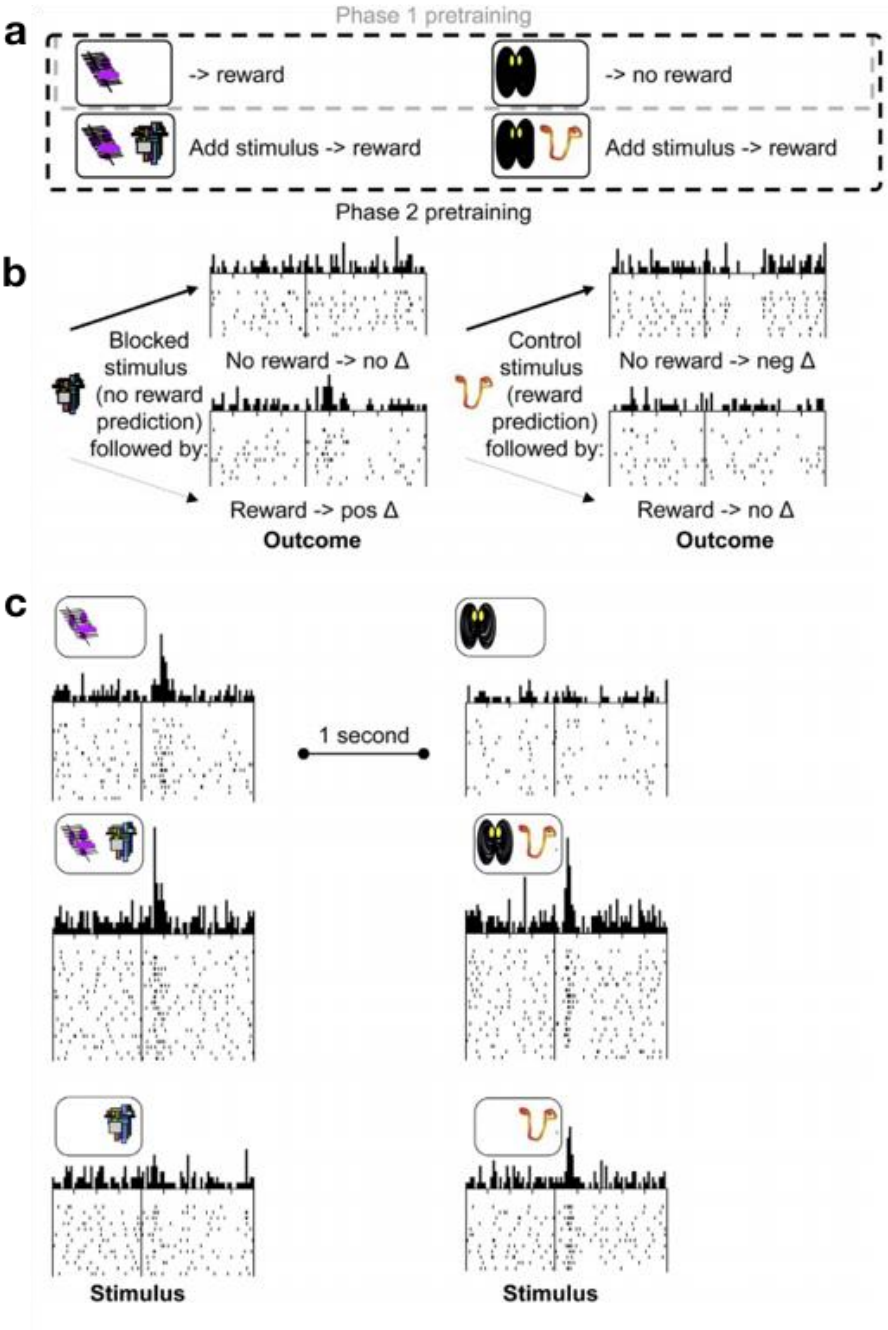


Figure 2.2. Dopamine neurons reflect the TD prediction error in a blocking task. Adapted from (Daw and Tobler, 2014), who in turn adapted from (Waelti et al., 2001). **a.** Task design. **b.** Response of a dopamine neuron at outcome time. The blocked stimulus does not predict reward, and so maintains its response when no reward is given (top left) or shows a positive prediction error when reward is given (bottom left, Δ is the TD prediction error). The contrary is true for the control stimulus, which does predict reward (right). **c.** Responses of a dopamine neuron at stimulus presentation time. The responses follow the predictions of the TD model (section 1.3.3)

The primary targets of DA neurons are the striatum and the prefrontal cortex. Indeed, extracellular dopamine levels (measured through fast-scan cyclic voltammetry) in the vStr also reflect the TD RPE, matching the electrophysiological data from the midbrain (Day et

al., 2007). However, in recent years there has been a surge of evidence from rodent studies, including from such fast-scan cyclic voltammetry, that the Dopamine signal does not necessarily code a "pure" RPE signal (and might even code for completely different variables like distance to reward (Howe et al., 2013)). For example, in a Go/noGo task where cues instruct an animal to either initiate or abstain from movements to obtain (an equal amount of) reward, there was RPE coding in the DA concentration levels in ventral striatum (vStr) only for correct action initiation following the cues that require movement initiation. In noGo trials, the cue did not elicit an RPE DA response. The DA response was delayed until the rat initiated a movement after successfully waiting for the required amount of time (Syed et al., 2016). Thus, mesolimbic DA concentrated are modulated by correct action initiation, not only RPEs.

Similarly, there is an accumulation of evidence from recordings of DA neurons that question the "pure RPE" hypothesis (Langdon et al., 2018). For example, DA neurons signals errors in the prediction of reward identity, even when the value of the rewards is controlled for (Takahashi et al., 2017). An important reason that made the RPE theory of dopaminergic firing attractive is the relative homogeneity of DA neurons (Cohen et al., 2012; Ungless et al., 2004), as is required from a "broadcast" learning signal as the RPE. However, several studies have now questioned this homogeneity (Parker et al., 2016), even suggesting DA neurons are functionally and anatomically organised in different clusters coding for different sensory, motor and cognitive variables (Engelhard et al., 2019). In addition, recent elegant work showed that dopamine transients are both necessary and sufficient for a rat to learn stimulus-stimulus association (i.e. a transition structure map) of the environment, in the absence of any external reward (Sharpe et al., 2017).

It's worthwhile noting that the connection between dopamine and prediction error seems to be restricted to rewards, and does not apply to efforts – even when controlling for the weight of reward and effort on behaviour (Walton and Bouret, 2019).

2.1.1 Prediction error coding in fMRI

In an apparent mirroring of the animal literature, BOLD signal in the vStr correlates with the RPE (appetitive rewards: (O'Doherty et al., 2003) aversive outcomes, as well as evidence for TD PEs in high-order conditioning): (Seymour et al., 2004); weaker response to blocked than unblocked CSs: (Tobler et al., 2006). For a meta-analysis see (Chase et al., 2015)). This signal has been demonstrated to be causally linked to DA levels, and to influence behaviour

in the directions predicted by TD algorithms (Pessiglione et al., 2006), mirroring demonstrations of the causal influence of dopamine on behaviour in animals (Steinberg et al., 2013).

Historically, this has implicitly been taken as a reflection of the DA input to vStr rather than the activity of vStr neurons per se (O'Doherty et al., 2004; Pessiglione et al., 2006). However, other accounts suggest that when reward is not the important dimension the subject needs to track, the vStr BOLD signal reflects the prediction error about variables that are relevant to behaviour rather than reward, unlike VTA BOLD signal which reflects the classic RPE even in Pavlovian tasks where the subject does not need to track - as predicted by the TD model (Klein-Flügge et al., 2011). Note that other account of the VTA BOLD signal have suggested it carries other information than classic RPE, for example a reward identity prediction error on value-matched rewards (Takahashi et al., 2017) that is independent from the rewards perceptual features (Suarez et al., 2019).

Prediction error signals in fMRI are not restricted to the vStr. In particular, another major target of midbrain dopamine neurons (the mesocortical pathway) is the vmPFC. Prediction error signals in this region many times accompany the stronger (and hence more popularly reported) vStr prediction error signal (Daw et al., 2011; Hampton et al., 2006; Ramnani et al., 2004; Wimmer et al., 2012). Indeed, stimulating the VTA, direct perfusion with dopamine, and the application of a DA reuptake inhibitor all lead to a sustained increase in long term potentiation (LTP) in PFC (Garris et al., 2006). This allows dopamine to influence cortical representations, supposedly by signalling the behavioural relevance (prediction error) of stimuli or actions.

Importantly, to date all of the fMRI prediction error studies have examined univariate signals of prediction error. In chapter 6 we will see that vStr and vmPFC prediction error signals can vary (in a fine-scale, multivariate manner) depending on *how PEs should be used* - even when they are computed in the same way and represent the same variable. This is different from most accounts of heterogeneity of dopamine coding in animals mentioned above, which usually refer to the coding of different variables.

2.1.1.1 Model-based and model-sensitive prediction errors

While the discovery of RPEs representations has transformed our understanding of the brain's model-free learning algorithms, there is still much that is unknown regarding the neural representations behind other types of learning algorithms. (Daw et al., 2011) used the 2-step task (section 1.3.5), where MF and MB learning and behaviour can be dissociated, to

investigate the PE representations in the vStr and vmPFC BOLD signal. They parallelly ran MF and MB algorithms which calculated different state values, and hence also produced different reward prediction errors. They then showed that the signal in vStr and vmPFC is better explained by including some contribution of the MB model, suggesting RPE-related signals in these areas are not purely MF.

Relatedly, (Gläscher et al., 2010) investigated a different kind of MB prediction error in a human, non-spatial version of Tolman's latent learning task. Here, participants first learned the transition structure of an MDP through observation, without any reward. This enabled the investigation of "state prediction errors" (SPEs), reporting discrepancies between the subject's current beliefs about the transition structure and the observed transitions. They later added reward contingencies at the final stage of the task, thereby enabling a comparison between SPEs and RPEs. Alongside the usual vStr RPE signal, they found SPE representations in the intraparietal sulcus and the lateral PFC.

(Hampton et al., 2006) used probabilistic reversal learning, a simple decision-making task with a higher order structure (anticorrelation between the reward probabilities of the two possible actions), to show RL signals follow the predictions of a model that uses the higher order structure of the task. Specifically, they show the vStr and vmPFC RPE signals are better explained by such a model than a model that separately learns about the values of the two possible actions. They refer to this as evidence that these regions perform model-based calculations. In section 1.3.6.3 I discussed how some behaviours that can seem MB can be obtained by applying MF algorithms on sophisticated state representations (Akam et al., 2015). I placed such strategies under the broad umbrella of "model-sensitive" algorithms. In chapter 6, I use this idea to interpret the results from (Hampton et al., 2006), and my own replication of their results.

2.2 **Evidence for multiple dissociated learning systems, and action representations in dorsal striatum**

This thesis does not include any novel results or insights regarding the dorsal striatum (DS). However, it is a crucial part of the action-selection circuit in the brain, and it is the source of some of the cleanest dissociations between habitual and goal-directed behavioural systems, a theme that is relevant for this thesis.

2.2.1 Evidence for parallel model-free and model-sensitive learning systems

Early evidence for the existence of parallel and separate learning systems came from elegant lesion studies such as (Packard and McGaugh, 1996) who revived one of Tolman's experimental design to demonstrate that actually, both he and the S-R behaviourists were correct regarding the brain's learning mechanisms. (Packard and McGaugh, 1996) showed a dissociation between the dorsolateral striatum (DLS, here referring to the caudate nucleus) and the dorsal hippocampus (HPC) in the learning of stimulus-response (S-R) and stimulus-stimulus (S-S, in other words, map-like) learning, referred to in the task as "response vs place" learning, respectively. Rodents were extensively trained on an elevated plus maze. At the beginning of each trial, they were placed at the start of the same arm ("south"), and when reaching the junction, had to turn right to obtain the reward, which was always placed at the end "east" arm (**Figure 2.3**a left). After both 8 and 16 days, animals were presented with the critical probe trial: placed at the opposite arm ("north" to the usual starting arm, and without any reason to believe the reward has changed location, will they perform the action that has been rewarding in the past (turn right, "response") as is predicted by S-R theories, or will they run to the old location of the reward (turn left to the east arm, "place") even though that action has never been rewarded before (**Figure 2.3**a right)?

After 8 days of learning, normal rats displayed "place" choices, indicating they could use a map of the environment to select even non-reinforced actions. However, after 16 days, most normal animals displayed "response" choices, consistent with previous accounts of a transfer to S-R, habitual behaviour with extensive training. Critically, inactivation of DLS and HPC just prior to the probe test had dissociable effects on the animals. Inactivation of HPC abolished the ability to use the cognitive map to guide behaviour on day 8: "response" behaviour was chosen as frequently as "place" behaviour. Inactivation of DLS abolished the transfer to habitual behaviour on day 16: animals kept showing "place" behaviour (**Figure 2.3**b). Hence, place and response strategies are mediated by dissociable brain regions – HPC and DLS, respectively. Though in normal animals a shift to a DLS-dependent response strategy is observed with extensive training, the HPC-based place strategy remains intact and can be uncovered.

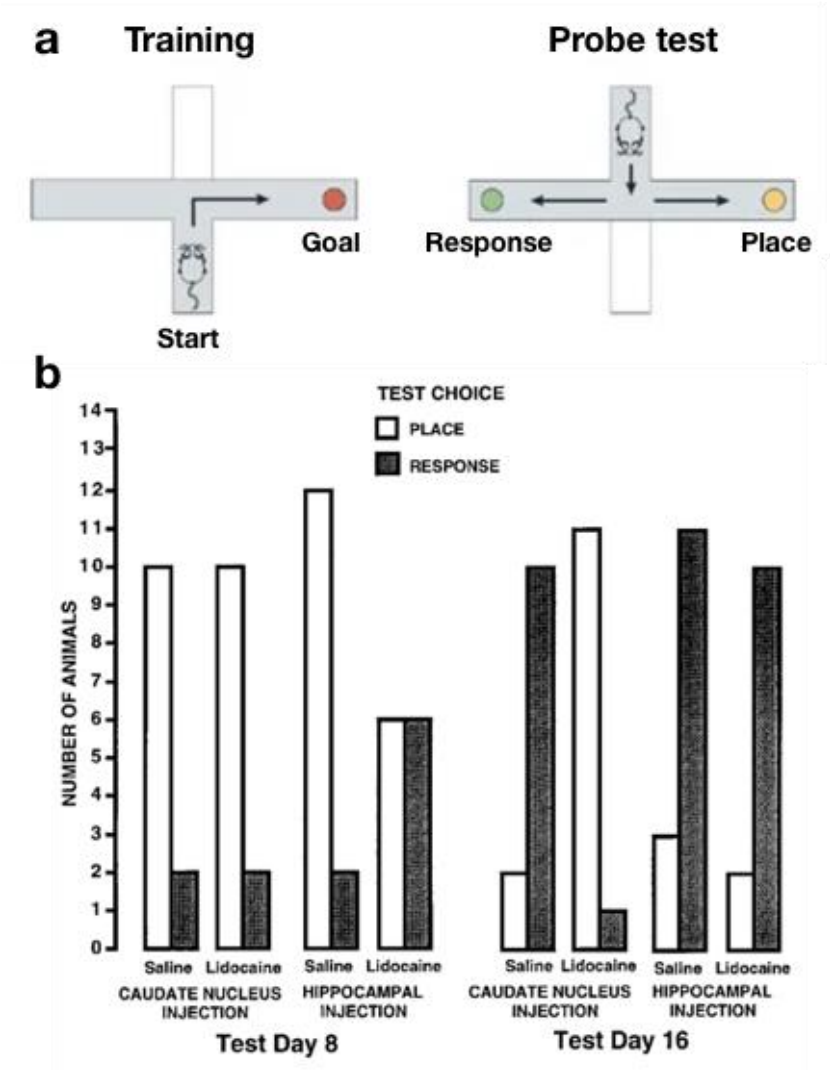


Figure 2.3 Parallel and dissociated learning systems for S-R ("response") and map-like ("place") learning. a. Task setup. Adapted from (Yin and Knowlton, 2006). **b.** Dissociable results of HPC and DLS inactivation. Results are described in the main text. Adapted from (Packard and McGaugh, 1996).

Evidence for a dissociation between the habitual and goal-directed neural systems comes also from non-spatial tasks, such as the devaluation tasks. Particularly well-studied is the dissociation between the sensorimotor, lateral parts and associative, medial parts of the dorsal striatum (DLS and DMS, respectively), where the former is selectively important for habitual behaviour, and the latter for goal-directed behaviour (Balleine, 2005; Yin and Knowlton, 2006). Inactivating DMS either before or after training made animals insensitive to devaluation, indicating a deficit in goal-directed action-outcome learning (Yin et al., 2005). On the other hand, animals with DLS lesions reduced their responses to a devalued outcome like control animals, but did not show any residual habitual behaviour like control animals

(as indexed by still pressing the lever associated with the devalued outcome in an extinction test (Yin et al., 2004)).

Similar dissociation can be observed when examining and comparing the representations in DLS and DMS. (Thorn et al., 2010) recorded from both areas while animals performed a cued T-maze task, where a sensory cue informs the animal of the left/right choice they need to make. DLS developed ensemble activity that was heightened at the action boundaries of the task (at the start of a trial or during turning to one of the arms). This representation steadily increased during learning and correlated with performance. In contrast, DMS was mostly active after the cue - i.e. when the choice had to be made. This activity pattern was not there at the beginning of training (when animals still did not understand what the task was about), developed (when animals understood the task, and supposedly used prospective planning to solve it), and then disappeared with over training (as the task became habitual with over-training).

These results, together with many others, have led to suggestions that the DLS represents something akin to the policy in the "actor" from the actor-critic MF model **Figure 2.4**: a representation of the choice history and the current task's behavioural repertoire (Sharpe et al., 2019). Meanwhile, the DMS codes for much richer information about the task (akin to a state representation), as is expected from a region involved in goal-directed behaviour. (Stalnaker et al., 2016). Suitably, this representation is dependent on the OFC input – a region suggested to code for the current state of the task (section 2.6.4, (Schuck et al., 2016; Wilson et al., 2014)).

It is noteworthy that both regions, as well as the vStr, receive dopaminergic input, suggestively enabling the training of the DLS actor with the critic's prediction errors (**Figure 2.4**). However, the topographical organisation of the dopaminergic input to striatum casts doubt over the notion of a unitary "broadcast" dopaminergic prediction error signal to both DLS and VS.

The dissociation between the habitual and goal-directed brain systems also has parallels in frontal cortex, with cross-region connectivity showing consistency with the notion of two dissociated systems. Prelimbic lesions abolish sensitivity to devaluation (Balleine and Dickinson, 1998; Corbit and Balleine, 2003; Killcross and Coutureau, 2003). In contrast, the transfer to habitual behaviour following extensive training is attenuated following lesions to the infra-limbic region (Killcross and Coutureau, 2003). Human fMRI studies have indicated a parallel division: the anterior caudate together with mPFC and mOFC are important for

goal-directed behaviour, while the posterior lateral putamen is critical for habitual learning (Balleine and O'Doherty, 2010).

2.3 **Value representations in the ventral striatum**

The strong reciprocal connections between the VTA and the vStr mean that prominent accounts of vStr function are closely related to those of dopaminergic function. I start by discussing a suggested role of this circuit in the brain's habitual learning system, as there are elegant mappings between theory and data to be made. However, it is important to note the vStr plays a critical role in representing values in Pavlovian scenarios too (such as the blocking experiment presented above (McDannald et al., 2011)).

As detailed in section 1.3.3.1.3, in the actor-critic model the role of the prediction errors within the "critic" is to train the state-values. The ventral striatum (vStr, usually meaning the nucleus accumbens and the ventral caudate/putamen) has been proposed to harbour these representations, together with parts of the PFC (section 2.5) (see (van der Meer and Redish, 2011) for a review and integration with other proposed functions of the VS like mediating motivational influences on decision-making during decisions, rather than a purely "evaluative" role). A prominent view is that this "training" occurs through the modulatory control of DA over cortico-striatal synapses: for example in a self-stimulation paradigm, where rats learned to perform an action to receive a stimulation of DA neurons, DA release in striatum induced potentiation of cortical-striatal synapses. The degree of potentiation correlated with how long it took to the rats to learn to self-stimulate themselves. (Reynolds et al., 2001). This gives a biologically plausible mechanism to a neural implementation of important components of the actor/critic model.

Neurons in vStr exhibit a strong response following the administration of reward (Carelli, 2002; van der Meer and Redish, 2009), and show ramping activity when approaching rewards (Lavoie and Mizumori, 1994), as would be expected from TD value representations. Consistent with this view, vStr is necessary for learning in a habitual learning task, but not for choosing actions correctly after learning has already occurred (i.e. after the actor is already trained (Atallah et al., 2007)), and is crucial for a range of MF tasks like second-order conditioning (where the conditioning is not sensitive to the current value of the reward, (Setlow et al., 2002)), conditioned place preference (Everitt et al., 1991) and outcome identity-general Pavlovian to instrumental transfer (Corbit et al., 2001; Corbit and Balleine, 2011).

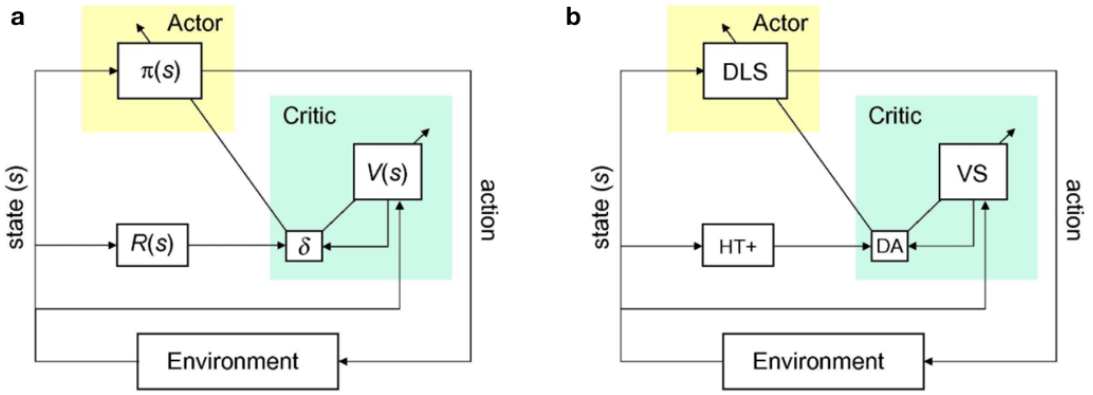


Figure 2.4 suggested mapping of the actor-critic model to structures of the basal ganglia. DA: Midbrain dopaminergic neurons (VTA/SNpc); VS: ventral striatum (vStr in the text). DLS: dorsolateral striatum. HT+: hypothalamus and other visceral structures. Adapted from (Botvinick et al., 2009), for a review see (Joel et al., 2002).

In the last decade it has become increasingly clear that the VS is also important for MB computation, as is demonstrated by finding that VS BOLD prediction error effects also reflect MB PEs (Daw et al., 2011). VS is critical for goal-directed representations of specific reward identities, as indexed by devaluation (Singh et al., 2010) and unblocking (McDannald et al., 2011) paradigms. In fact, a potential MF-MB double dissociation within the nucleus accumbens (NA) has been observed in a Pavlovian to instrumental transfer task (PIT), whereby the NA core is crucial for general but no outcome identity-specific PIT, while the NA shell is crucial for outcome-specific PIT, but not for general, "common currency" PIT (Corbit and Balleine, 2011). VS also takes part in tasks which require prospective planning, where it represents future reward in important choice points (van der Meer and Redish, 2009), Taken together, it seems the VS is a crucial node in the reward learning system for both MF and MB learning strategies.

2.3.1 The role of the amygdala in calculating expected value representations.

The amygdala projects to both VS and VTA and is critical for Pavlovian stimulus-outcome association. Some have even suggested the vStr value representations might be inherited from value representations in the amygdala (Averbeck and Costa, 2017), and that hence vStr does not actually play the central role in learning stimulus-outcome association as is widely believed. The amygdala is also widely activated in fMRI RL studies, for example coding for RPEs in Pavlovian tasks (Chase et al., 2015). It is thus obvious it is a crucial node

in the reward network of the brain. However, an in-depth discussion of the amygdala's role in this circuit is beyond the scope of this thesis. For a review, see (Averbeck and Costa, 2017).

2.4 **Model-sensitive representations in the hippocampus**

In the next chapter, I thoroughly discuss the extensive hippocampal formation literature suggesting it has a crucial role in the representation of cognitive maps suggested by Tolman – in both spatial and non-spatial tasks. For completion, I very briefly mention here several hippocampal phenomena which are useful for consideration within the current chapter's topic of RL representations.

In spatial tasks, a substantial percentage of hippocampal cells code for particular locations in the environment – a discovery that has led to the suggestion the hippocampus encodes Tolman's cognitive map ((O'Keefe and Nadel, 1978), section 3.1.1). However, hippocampal cells actually encode much more than just the current spatial location of the animal. For example, at important decision points, hippocampal representations in the theta frequency band will "sweep forward" to encode future possible locations (Johnson and Redish, 2007). Moreover, during rest or sleep, such sweeps (on an even quicker timescale - "sharp-wave ripples"), which strongly predict future behaviour (Pfeiffer and Foster, 2013), can follow trajectories to parts of the environment that have not been experienced yet – a phenomenon known as preplay (Dragoi and Tonegawa, 2011). Relatedly, human patients with hippocampal lesions are not able to imagine future states (Hassabis et al., 2007) and rats with HPC lesions show less VTE (vicarious trial and error, section 1.2.2) behaviour (Hu and Amsel, 1995). Such future "simulations" during planning are also the hallmark of model-based learning algorithms (section 1.3.4).

Recently, several studies have also confirmed the importance of the hippocampus for model-based behaviour in non-spatial tasks. As discussed in sections 1.3.5 and 2.1.1.1, the 2-step task (Daw et al., 2011) is the task usually used to dissociate MF and MB contributions to behaviour. Human patients with hippocampal lesions indeed rely more on MF strategies in this task (Vikbladh et al., 2019). Correspondingly, in an adaptation of the task for rats, inactivation of the dorsal hippocampus also causes an impairment in model-based planning, while leaving other strategies intact (Miller et al., 2017).

Importantly, human studies suggest the hippocampus can support apparently model-sensitive behaviour in other ways than simulation or prospective representation at decision time, through relational associative memory and generalisation (see next chapter for details) during initial learning (Shohamy and Wagner, 2008; Wimmer and Shohamy, 2012; Zeithamova et al., 2012). For example, (Wimmer and Shohamy, 2012) exposed subjects to pairs of stimuli (e.g. A&B,C&D, similar to a sensory pre-conditioning paradigm), prompting the formation of unconscious (A-B, C-D) associations. Some of the stimuli (A, but not C) were later Pavlovianly rewarded. When asked to choose between stimuli that were never rewarded (choose either B or D), participants had a bias towards stimuli associated with rewarded stimuli (B), as is expected in a sensory pre-conditioning paradigm. The amount of this bias was correlated with the hippocampal activity during the reward time (of A), as well as with the functional connectivity between the hippocampus and a reward-sensitive area of the striatum – suggesting that hippocampal associations helped to "spread" the reward to the associated, non-rewarded stimulus (A), even in the absence of any conscious associations (Wimmer and Shohamy, 2012).

(Stachenfeld et al., 2017) suggested the hippocampus reflect a different (yet very related) model-sensitive entity – the successor representation (section 1.3.6.1). I cover this suggestion, more thoroughly in section 3.2.3.1, including some remarkable fits of the successor representation to hippocampal neural signals in discrete abstract maps (Garvert et al., 2017; Schapiro et al., 2013), suggesting the hippocampus does indeed code a predictive map of abstract spaces.

2.5 **Value representations in vmPFC/mOFC (BA 14)**

The human vmPFC is perhaps the most intensely studied PFC region in the context of reward-guided learning. A meta-analysis of fMRI studies found that together with the ventral striatum (and to a lesser extent the posterior parietal cortex), the region where positive effects of subjective value of choices have been found most consistently is the vmPFC (Bartra et al., 2013; Chase et al., 2015). The vast majority of these effects centre on the medial surface of the frontal lobe extending to the medial parts of the OFC, in an area which a connectivity study based on diffusion-weighted imaging identified as a having specific connectivity patterns (green arrow in **Figure 2.5**b). This area has particular cytoarchitectonic properties according to (Mackey and Petrides, 2010) who refer to it as BA 14m (**Figure 2.5**a).

Importantly, a proposed homologous area to human BA 14m in the macaque is the medial orbital gyrus (more ventral than in humans) – where some of the reward-related macaque signals have been recorded (Strait et al., 2014).

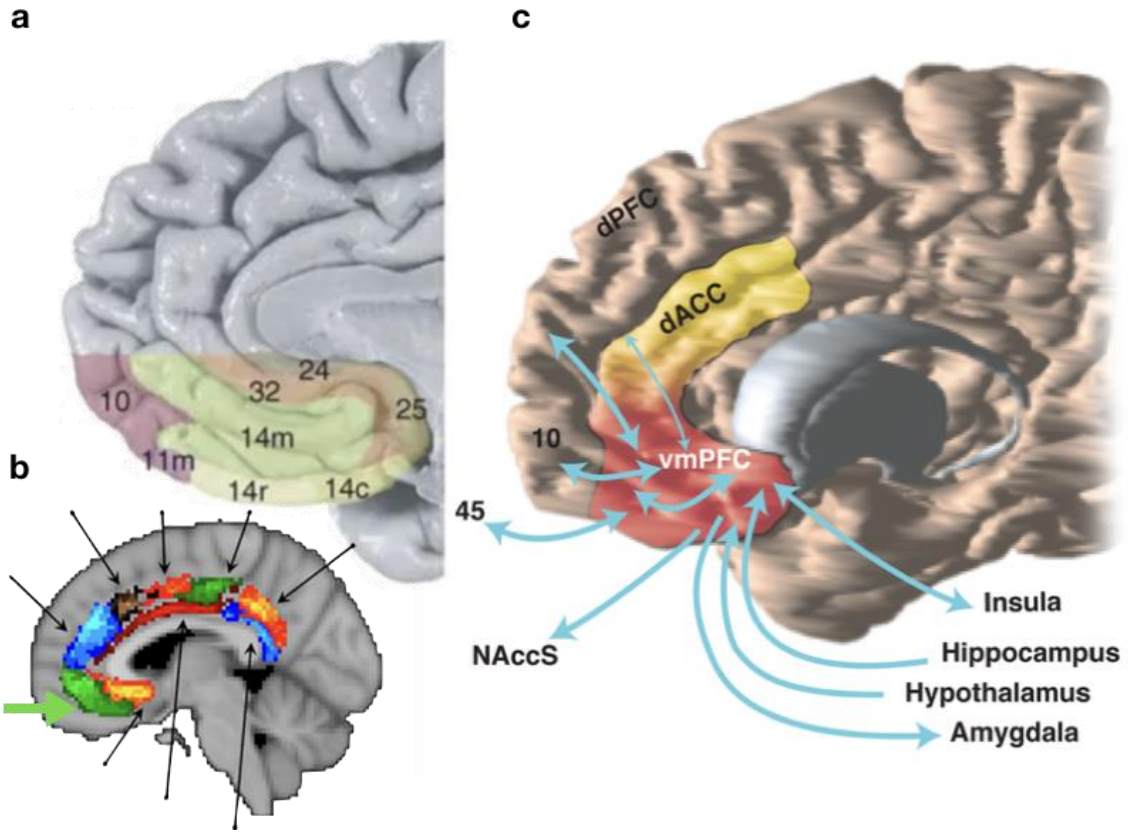


Figure 2.5 The ventromedial PFC. a. Adapted from (Mackey and Petrides, 2010). **b.** For our purposes, black arrows should be ignored. Adapted from (Beckmann et al., 2009). **c.** Schematic of the information flow to and from vmPFC. From (Haber and Behrens, 2014).

There are several, not mutually exclusive, accounts of the type of value signals this area harbours. The BOLD signal in this area correlates with the subjective value participants assign to single food items as indexed by their willingness to pay for their consumption (Plassmann et al., 2007). This is also true for the aversive analogue: the vmPFC will be less active during presentations of the more aversive food items, as indexed by subject's willingness to pay to avoid their consumption (Plassmann et al., 2010).

The vmPFC signals both the value of experienced reward (at outcome time - (Smith et al., 2010; Strait et al., 2014), a signal which is very close and correlated (yet dissociable with the right experimental design) with the vmPFC prediction error signals discussed in section 2.1.1), and the value of different options at decision time. One of the most ubiquitous vmPFC findings is the signalling of the difference in value of different options (Boorman et

al., 2013, 2009; Juechems et al., 2017; Kolling et al., 2012). Specifically, it seems to reflect a behaviourally useful decision variable: [the value of the chosen option] – [value of the unchosen option] ((Boorman et al., 2009), **Figure 2.6**). These results mirror evidence from lesion studies, suggesting the vmPFC is needed to make simple value comparisons. Patients with vmPFC lesions defy transitivity axioms when choice is based on subjective value, but not in a control test where they choose according to perceptual differences (Camille et al., 2011; Fellows and Farah, 2007). Perhaps even more convincingly demonstrating the importance of this regions to value comparison is the finding that macaques with mOFC lesions show a deficit of choosing between options with different values that increases as a function of the proximity of the choices' values (Noonan et al., 2010).

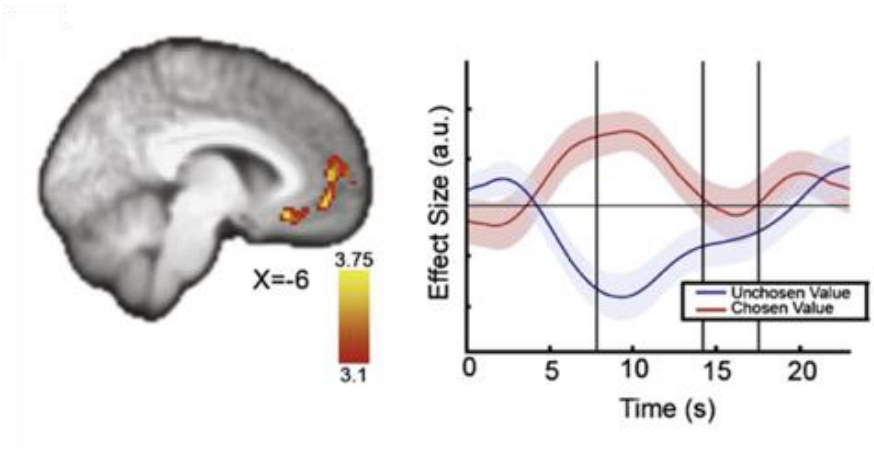


Figure 2.6 Value comparison signals in the vmPFC. Left: Effect of the contrast [Value of Chosen option] – [Value of Unchosen option]. Right: Time course of the two components of this effect from a vmPFC ROI. From (Boorman et al., 2009).

However, not all studies have reported a vmPFC value difference signal, and some report different types of value signals. (Wunderlich et al., 2010) found that the values of both options were represented early in the trial, but only the value of chosen option is represented late – without any representation of value difference, suggesting the comparison between the options was performed by a different mechanism. (Juechems et al., 2017) showed that during outcome time, alongside the anticipated signal coding for the received reward, there was a signal tracking the cumulative reward that has been obtained so far (within an arbitrary "section" of the task, with the division into sections being irrelevant for behaviour). This was true even though there was no incentive for tracking this quantity. Different value signals can also be present at different time points in the trial, inaccessible to sluggish fMRI: (Hunt et al., 2012) used MEG to show that both the sum of the values of options and the difference

between them are represented in vmPFC – with the former appearing before the latter, suggesting vmPFC is involved in both valuation and selection. Other authors have not found any vmPFC value signals (Kolling et al., 2014). Taken together, it seems that vmPFC value signals are paradigm dependent.

Why should the brain encode value signals in multiple brain regions? Indeed, the differences value signals in vmPFC and its target vStr are still not clear. Some studies have suggested these regions have different functions, where the vStr is usually linked to more automatic action selection and learning, while vmPFC is more associated with flexible, model-based behaviour. As we saw in section 2.3, this division should be considered as very loose and flexible. Indeed, other studies have suggested both regions code for very similar information (Strait et al., 2015).

An influential study that made the connection between vmPFC value signals and model-based behaviour was (Hampton et al., 2006), which we have already discussed in the context of prediction errors in section 2.1.1.1. In this reversal learning task, the authors showed that the value signal in vmPFC is better explained by a model that knows about the task structure (anticorrelation between reward probabilities of the two options) than a "naïve" model that learns about each option separately. In chapter 6 I replicate these results, but suggest a different interpretation, which is model-sensitive in nature but not "pure" model-based: a model-free algorithm, operating on a sophisticated task space that knows about the latent structure in the task (Akam et al., 2015). Indeed, it seems both MF and MB value signals can be found in vmPFC (Beierholm et al., 2011).

A recent study has suggested a different perspective on the difference between vStr and vmPFC reward signals. When the task is not only to maximise reward but also maintain equilibrium between different dimensions of reward, vStr code the reward regardless of whether it fits current needs (i.e. regardless of the distance from equilibrium), while vmPFC codes how valuable the outcome was to restore equilibrium ("redress", (Juechems et al., 2019)). The authors then suggest a more general framework for reward-guided learning, where rather than trying to maximise external rewards, agents seek to minimise the distance to goals along several dimensions (Juechems and Summerfield, 2019).

The suggestion that the vmPFC integrates information along several dimensions relevant to goals and reward point to a feature of vmPFC representations which is of particular interest to this thesis. In the next chapter, we will discuss the role of the vmPFC, through its functional connections with the hippocampal formation, in tasks where *relational* knowledge

is important. This is also true for vmPFC value signals, and can perhaps offer some explanations to the discrepant, paradigm-dependent results regarding vmPFC value signals discussed above. For example when subjects evaluate novel food items armed with knowledge about their familiar components, it is the vmPFC and hippocampus that construct the novel experience (Barron et al., 2013). Relatedly, patients with vmPFC lesions are selectively impaired in "configural" (i.e. relational) valuation: subjects were exposed objects ("Freebles") built out of several elemental building blocks, with only two elements varying. They learned arbitrary values of these elements by viewing "Freebles prices in an auction", and were then tested on knowledge of these values until a learning criteria was reached. Crucially, in one condition, the value of the objects was determined by the sum of the two of value-predicting elements ("elemental"), while in the other condition each pair of varying elements (i.e. object) had a unique "configural" value, without any useful information about the value of individual elements. Patients with vmPFC lesions (unlike patients with vmPFC-sparing frontal lesions) were specifically impaired in valuating objects where such a "configural valuation" was needed, but performed as controls for objects without any of these "special pairs". That is, when the object value could be calculated based on the elemental values (even when two elemental attributes differed between the objects two), vmPFC patients performed normally. (Pelletier and Fellows, 2019). Note, however, that when the authors analysed decisions on a subset of trials with the most similar attribute processing and value-difference requirements across conditions (elemental trials where not only the two value-predicting trials differed between the options (namely two-attribute trials), but both wre also necessary for correct choice; configural trials that value-difference matched matched these trials), they did not observe this effect – though this might be a result of low statistical power.

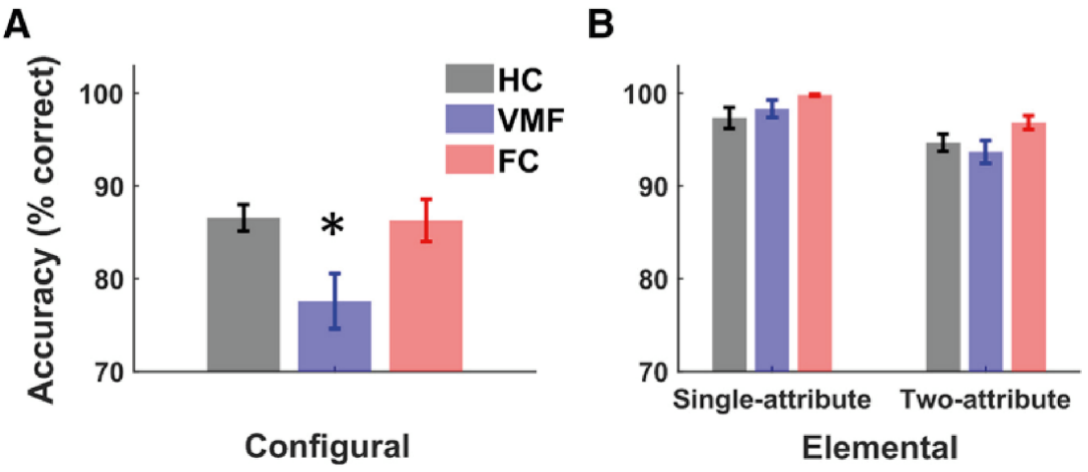


Figure 2.7 Patients with vmPFC lesions are selectively impaired in a valuation task where the relationship between elements is what matters. Lesion groups: VMF **-** ventromedial frontal cortex; HC - Hippocampus; FC – VMF sparing frontal lesions. In the configural condition subjects had to learn values that were unique to each pair of attributes (elements), while in the elemental condition values of objects depended directly on either a single attribute or the sum of two attributes. Note that the "two-attribute" trials include trials where even though 2 value-predicting attributes were different between the two objects, it was enough to take only one of them into account for correct choice. Adapted from (Pelletier and Fellows, 2019).

2.6 **State and learning set representations in OFC and vlPFC**

Perhaps more than any other brain region, the OFC has been implicated in virtually every aspect of cognitive tasks, with its involvement in some tasks being highly controversial (Stalnaker et al., 2015; Wallis, 2011).

Some of the complexity of the areas' function was illustratively communicated in a report about Phineas Gage, a railroad worker who was using an iron tamping rod to pack the explosive powder back into a hole when the powder detonated, sending the rod through Gage's frontal lobe, and in particular the OFC[1]. His doctor, John Harlow, described the dramatic post-accident changes in behaviour as follows:

> *"The equilibrium […] between his intellectual faculties and animal propensities, seems to have been destroyed. He is fitful, irreverent, indulging at times in the grossest profanity (which was not previously his custom), manifesting but little deference for his fellows, impatient of restraint or advice when it conflicts with his desires, at times pertinaciously obstinate, yet capricious and vacillating, devising many plans of future operation, which are no sooner arranged than they are abandoned in turn for others appearing more feasible. […] Previous to his injury, though untrained in the schools, he possessed a well-balanced mind, and was looked upon by those who knew him as a shrewd, smart business man, very energetic and persistent in executing all his plans of operation. In this regard his mind was radically changed, so decidedly that his friends and acquaintances said he was 'no longer Gage'* (Harlow, 1868).

[1] Modern reconstruction techniques have suggested Gage's lesion involved orbital regions such as unilateral BA 11 and 13 (OFC proper) and bilateral BA 12 (lateral OFC/vlPFC), as well as other frontal regions (BAs 8,9,10,24,32) (Damasio et al., 1994).

This account has prompted suggestions that the core function of the OFC is response inhibition, as suggested by a range of studies of reversal learning paradigms. However, in the last decade the suggestion that response inhibition is a core function of the OFC has been abandoned, after several, more nuanced experiment found contradicting evidence, which could also explain the deficits reported earlier (reviewed in (Stalnaker et al., 2015)).

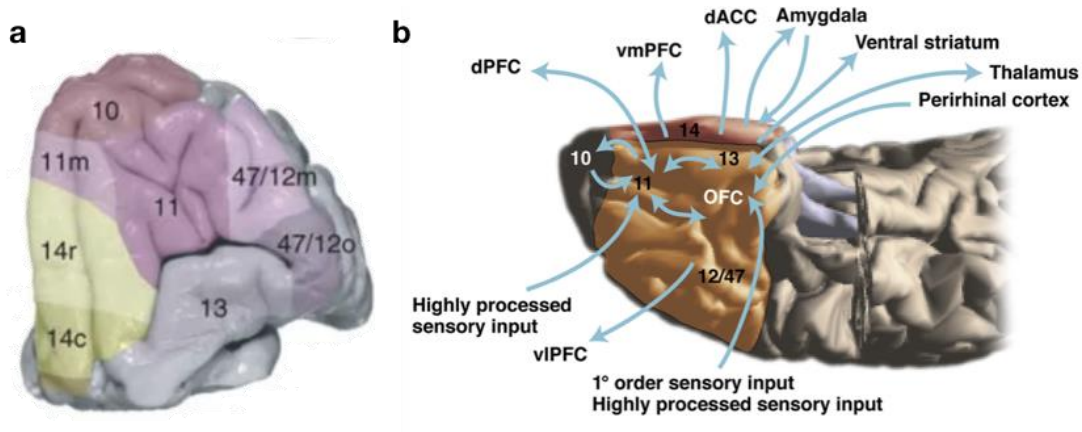


Figure 2.8 The orbitofrontal cortex. a. Cytoarchitectonic parcellation of human OFC. From (Mackey and Petrides, 2010). **b.** Schematic of information flow to and from the OFC. From (Haber and Behrens, 2014)

2.6.1 Central OFC (BA 11 & 13) encodes subjective value primarily in model-based tasks, perhaps mirroring human vmPFC BOLD signals

The standard view regarding OFC function is that its function is related to value-based decision-making. In rodents this function seems to be specific to model-based behaviour, as is demonstrated in the OFC-dependence of reward identity (but not value) unblocking tasks (McDannald et al., 2011). In monkeys and humans, the story is somewhat more complex. In all species, the OFC is important for devaluation effects (Pickens et al., 2005; Rudebeck et al., 2017, 2013), highlighting it's importance for goal-directed behaviour. Relatedly, in humans the caudal OFC represents reward identity independently from the stimuli predicting them (Klein-Flügge et al., 2013), and these representations are attenuated following devaluation (Howard and Kahnt, 2017), as is expected from a goal-directed valuation system.

In monkeys, the majority of value-based findings in the ventral PFC (spanning vlPFC, OFC and vmPFC) have been recorded in the central (sometimes referred to as lateral) OFC – BA 11 & 13. Some neurons encode specific information about an outcome ((Lara et al., 2009; Padoa-Schioppa and Assad, 2006), consistent with the aforementioned emphasis on goal-directed behaviour). On the other hand (and perhaps a finding that has been more influential), many neurons signal an abstract subjective value or "common currency" signal

(Padoa-Schioppa and Assad, 2006), which is invariant to the identity of the specific outcomes (juice), quantity, direction of required response, features of the predictive stimulus and even the period of the trial (this does not seem to be true in rats, where OFC value signals seem to be more strictly tied to reward identity (McDannald et al., 2011; Roesch et al., 2006)).

Such common currency signals, which predict idiosyncratic shifts in behavioural preferences and obey economic principles like transitivity and menu invariance[1] (Padoa-Schioppa, 2009; Padoa-Schioppa and Assad, 2008)), seem somewhat similar to the value and value difference signals found in human vmPFC in fMRI (e.g. (Levy and Glimcher, 2011)). However, the careful reader will remember that though a homology exists between human vmPFC/mOFC and monkey mOFC (BA 14 in both), there is no such homology between human vmPFC and monkey BAs 11 & 13 (which are homologous to human BAs 11 & 13). The possible reasons for this discrepancy - psychological, anatomical and technical, as well as many remaining questions about it, are reviewed in (Wallis, 2011). Nevertheless, in both monkey central OFC (Hunt et al., 2018) and lateral OFC in rodents, (**Figure 2.10** bottom, (Zhou et al., 2019)) representations are primarily organised based on value even when the task has other important dimensions, and these representations are used to guide choice as is evident by the decidability of choices and deliberation processes from OFC ensembles (Rich and Wallis, 2016).

While the theory of OFC (and vmPFC in humans) as the calculator of subjective value for decision making has been highly influential, it does not explain all findings. There are many value-guided choices for which the OFC is not necessary (Schoenbaum et al., 2011). For example, monkeys with aspiration lesions of areas 11, 13 and 14 showed devaluation deficits when choosing between objects that predicted different food items, some of which were devalued, but were unimpaired when requested to choose directly between the food items (Izquierdo et al., 2004). Together with an abundance of evidence from rodents (Gremel and Costa, 2013; Mcdannald et al., 2012; McDannald et al., 2011; Schoenbaum et al., 2011), such findings have led to a weaker, more nuanced value hypothesis of OFC: that it might be necessary only when behaviour is derived from a model-based mental simulation (Padoa-Schioppa, 2011). The critical feature of this hypothesis is an emphasis on processing of latent

[1] The independence of the value encoding of a particular good from alternative offered goods.

information – for example linking the (separately observed) stimulus-outcome and outcome-illness associations, even though they were not experienced together. Proponents of the economic value theory suggest that in all economic choices, values are actually computed on-the-fly and are not being just retrieved from stored cached valuations (Padoa-Schioppa, 2011). Under such a view, the strong and the weak versions of the value hypothesis converge.

2.6.2 OFC and vlPFC during learning

The OFC results discussed so far have all centred on its functions when decisions need to be made. For example, in the devaluation results from above (Izquierdo et al., 2004; Rudebeck et al., 2017, 2013), researchers inactivated the OFC before the probe test when subjects are asked to choose between devalued and non-devalued objects. However, the OFC is also crucial for learning under specific conditions, which many times resemble the decision-related OFC properties discussed above. For example, by elegantly using sensory preconditioning (section 1.2.3) to distinguish between cached (experienced) and inferred (unexperienced) value in combination with a blocking paradigm, (Jones et al., 2012) showed that OFC is needed for both value based decision and learning from value – but only when this value is inferred. Several other studies, reviewed in (Stalnaker et al., 2015), show similar results. Together with (inconsistent) reports regarding error signals in OFC, these results indeed point to the importance of the OFC for learning, perhaps (again) only in scenarios requiting mental simulation or with latent features (Stalnaker et al., 2015).

A critical, yet often overlooked part of learning problems is basic knowledge about how the task works. For example, a monkey needs to know that the reason he just got a reward was the stimulus he just chose, not the stimulus from 4 trials ago, and not winning the fight with his cage-mate an hour before. A deficit in this type of credit assignment problem was observed by (Walton et al., 2010) using an elegant 3-armed bandit design with carefully constructed probabilistic reward schedules (Figure 2.9**a**). Though (Noonan et al., 2010; Walton et al., 2010) originally suggested the deficit to be a result of lesions in areas 11 and 13, a later study (Rudebeck et al., 2017), using excitotoxic rather than aspiration lesions, concluded this deficit is actually a result of the more lateral vlPFC (BA 12, 45 and 46), not

OFC lesions[1]. Unlike previous reversal learning tasks, the usage of 3 stimuli enabled the authors to show that lesioned animals can track changes in reward, unless the best option switches (Figure 2.9**a**). Careful trial-by-trial analysis revealed the cause for this deficit – lesioned animals spread the credit for outcomes to past and future choices, instead of assigning it in a temporally precise manner (Figure 2.9**b**). Similar effects have been reported for rodents, where unilateral lesions of OFC cause dopaminergic neurons in the same hemisphere to report prediction errors based on a prediction that no longer reflects which choice has caused the reward (Takahashi et al., 2011). In addition, a human fMRI version of the task which contrasts activity following rewards that are contingent on the subject's choices to non-contingent rewards has highlighted bilateral regions of the lateral OFC as specifically encoding the contingent rewards ((Jocham et al., 2016), **Figure 2.9**c).

Some authors (e.g. (Bouret and Richmond, 2010)) have suggested a lateral-medial functional gradient in vPFC, where lateral areas code for values associated with external stimuli while medial regions for values associated with internal states (such as a deliberation between two options). This fits well with the localisation of stimulus-outcome association effect to the most lateral parts of the ventral PFC (i.e. vlPFC), and the suggestions that more medial parts are important for choice (in devaluation tasks (Rudebeck et al., 2017) or choices between two similar-valued options (Noonan et al., 2010)).

[1] (Rudebeck et al., 2017) suggested that the original results by (Walton et al., 2010) were probably a result of damage to fibre tracts between vlPFC and the temporal pole, that pass close to the OFC.

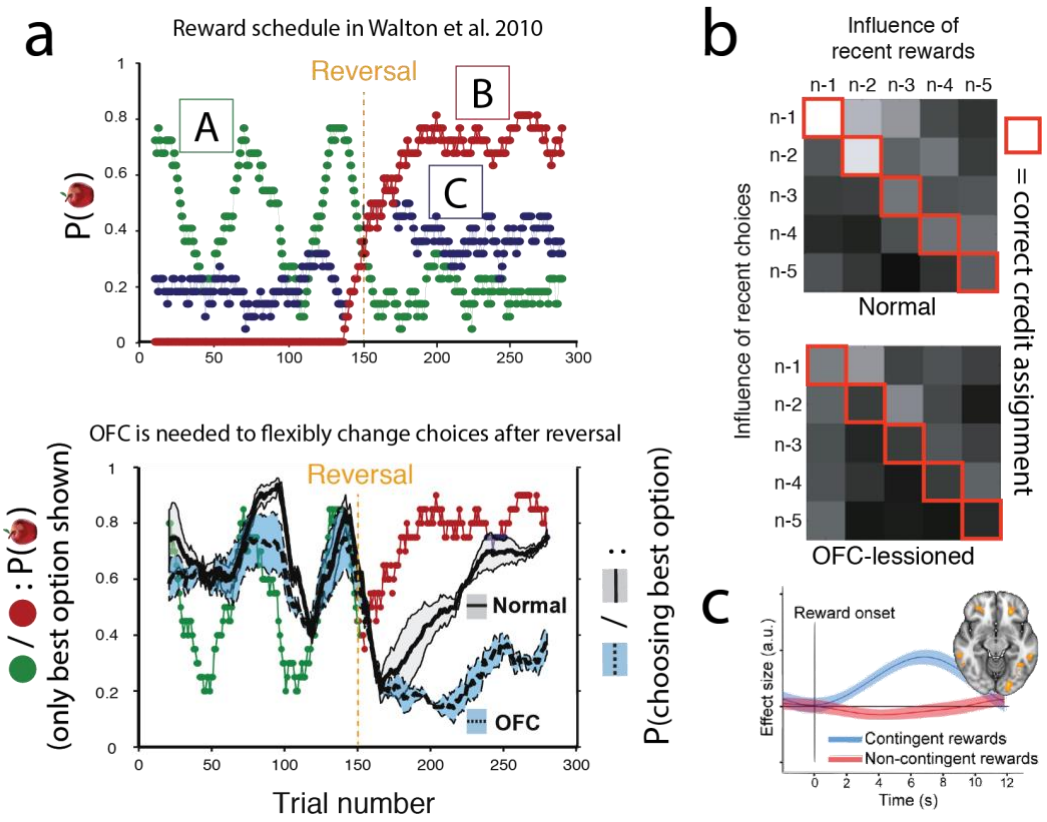


Figure 2.9 Credit assignment deficits in vlPFC-lesioned animals. a. vlPFC lesions animals can track changes in reward unless the best option switches. **b.** This deficit is because they no longer know which recent choice caused which recent reward. In these regression plots, correct attribution between choice and reward is on the diagonal. vlPFC-lesioned animals learn using the off-diagonal terms, which implies that if average recent reward is high (low) the animals increases (decreases) their preference for their average recent choice, ignoring which choice caused which reward. Greyscale colours are regression weights estimating the influence of recent choices and outcomes on current behaviour. From (Behrens et al., 2018b), where we adapted from (Walton et al., 2010). The localisation of this effect to vlPFC rather than lOFC was made by (Rudebeck et al., 2017). **c.** Human lOFC preferentially signals contingent rewards over non-contingent rewards. Brain maps showing whole-brain [contingent] – [non-contingent] contrast. BOLD signal at the peak of the lOFC effect is plotted. From (Jocham et al., 2016).

2.6.3 "The mysterious OFC"[1]

Why, then, has it been so difficult to describe the function of the OFC (Stalnaker et al., 2015)? At least four major reasons contribute to this difficulty. First, it is possible that the question is inherently wrong – OFC is comprised of several subregions, with different

[1] (Cavada and Schultz, 2000; Wallis, 2011)

connectivity patterns and functions, and hence trying to reconcile the functions of these subregions under a unifying theory is bound to fail. The highly confusing nomenclature of referring to different OFC regions is probably a significant contributor to this problem. Hence, I have tried to be as anatomically accurate and specific as possible when describing OFC results.

Second, as the OFC is anatomically close to fibre tracts, special care needs to be taken when interpreting aspiration lesion studies. For example, (Rudebeck et al., 2013) showed that the performance of OFC-lesioned in deterministic object reversal learning tasks (like the one Harry Harlow used, discussed in the introduction to the introduction) was completely on par with controls when the lesions were excitotoxic (damaging only grey matter), rather than aspiration lesions (which also damage passing white matter fibres). Until then, reversal learning (together with reinforcer devaluation) was considered the classic OFC-dependent task (in part due to cases like Gage's which were historically interpreted as a deficit in inhibitory control). Another example of this issue has been described above – see (Rudebeck et al., 2017).

Third, cross-species homologies of vPFC are controversial, and should be treated with care. Human OFC can be divided according to the presence of a granular layer (cortical layer IV), with virtually all vPFC areas except the most caudal parts of areas 13 and 14 including a granular layer to some extent ((Wise, 2008), **Figure 2.10**). Though there has been substantial extension of the granular layer from monkeys to humans, there is broad consensus that the structure and organisation of the human and monkey ventral PFC are conserved. Rodents, however, lack any granular vPFC cortex whatsoever. Despite these cytoarchitectonic differences, there are many anatomical and functional parallels between parts of the rodent brain (including the dorsal bank of the rhinal sulcus, lateral orbital regions and agranular parts of the insular cortex) and primates OFC (BA 11,12 and 13 (Preuss, 1995; Schoenbaum et al., 2009), **Figure 2.10** bottom. Notice the exclusion of area 14 from this figure – there is no suggested homology for primate vmPFC in rodents). Together with remarkable functional similarity across species following lesions to these areas, there is a good reason why many authors still see the study of the rodent OFC as a very useful contribution to our understanding it in humans.

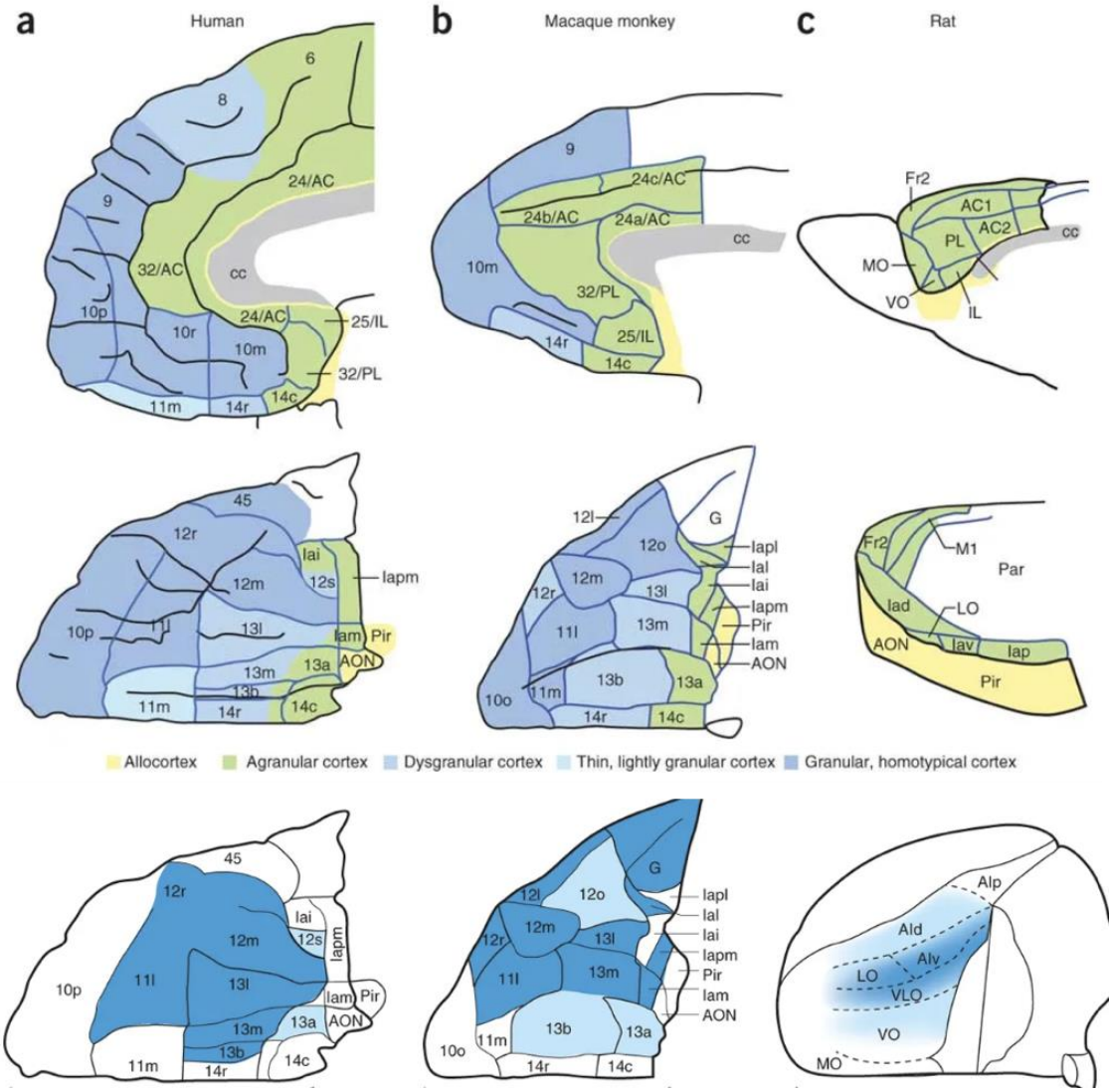


Figure 2.10 Differences and similarities between primate and rodent OFC. Top and middle panels: **a-b**. Architectonic maps of the medial (top) and orbital (middle) surfaces of the frontal lobe in humans (a.) and monkeys (b.). **c.** Medial (top) and lateral (middle) frontal cortex in rats. Importantly, the division to areas in vmPFC has been revised from the time this figure was produced, and it is now considered to be part of area 14 in both humans and monkeys, indicating a strong homology ((Mackey and Petrides, 2010), Figure 2.5). Agranular cortex lacks layer IV. Dysgranular cortex contains a rudimentary layer IV. Granular cortex has a well-developed layer IV. Layer IV neurons are described as granular because their cell bodies are small and round. The transition from agranular to granular cortex is accompanied by a clearly visible change in this layer. From (Wallis, 2011) who in turn adapted from (Wise, 2008). Bottom panels: suggested correspondence of the lateral (dark blue) and intermediate (i.e. in both lateral and medial networks, light blue) networks of the central OFC across species (notice the exclusion of the mOFC and vmPFC from this figure, for which there is no homologous rodent area suggested). From (Stalnaker et al., 2015).

2.6.4 State representations, task structure and learning sets

The fourth possible reason for the difficulty in describing OFC function might be a lack of the right general theoretical framework – and it is not actually a reason, but rather a recent

suggestion. The general consensus that there is some value-related coding happening in OFC and the success of formal RL theories in explaining other neural signals in the reward system have led to attempts to provide such an overarching theory for OFC function. Given the dominance of the economic value theory, in RL terms this usually implied that the OFC represents the value of a state – perhaps specifically in model-based or complicated hierarchical state spaces (Botvinick et al., 2009). However, as was discussed in section 1.3.6.3, most of RL theory has focused on learning algorithms, applied on a *given* state-space – without asking where the state-space came from, and how is it being represented.

In an influential paper, (Wilson et al., 2014) hypothesised that the OFC represents the current state (location within the task space), particularly when the information needed for the state definition is latent. While on first glance this hypothesis might seem very general, it generates remarkably precise predictions. By modelling control agents with the correct state-spaces and "OFC-lesioned" agents for which separate latent states collapse together based only on their observable features, (Wilson et al., 2014) were able to explain a wide range of results from the OFC literature, including deficits in reversal learning, devaluation, delayed alternation, extinction and precise credit assignment (Takahashi et al., 2011). However, it is important to note that though (Wilson et al., 2014) refer to the different experiments they model as "OFC" experiments, these data were actually obtained from manipulations or recordings in brain regions that differ markedly. For example, data from the Schoenbaum lab (e.g. (Takahashi et al., 2011)) is usually recorded from rodent lateral OFC and sometimes even anterior insula – whereas data from other experiments modelled in (Wilson et al., 2014) like devaluation effects in (Pickens et al., 2003) include lesions in medial OFC – areas that are very different both functionally and anatomically (Izquierdo, 2017)fa. Nevertheless, the explanatory power and possibility to generate clear hypotheses afforded by such a (perhaps simplified) view of a brain region have no doubt proved useful.

(Schuck et al., 2016) directly tested the hypothesis of (Wilson et al., 2014) with fMRI in humans, by developing a task where decisions were dependent on knowledge about the current phase (state) of the task, information about which was latent (involving memory of past events). Subjects had to judge the age (old vs young) of a face or a house (**Figure 2.11**a). In each trial both house and face were displayed, superimposed on each other. The rule determining whether the age of the face or the house should be reported depended on previous (currently unobservable) trials: as long as the age of the current category stayed the same, subjects should continue reporting the age of that category. Once they encounter a

trial in which the age of the currently judged category was different, they should switch a category. The true state-space of this task consists of 16 states, even though the physical appearance of each trial is very similar (**Figure 2.11**b). The current state could be decoded well above chance from the OFC (anatomically defined, including BAs 11,13,14, **Figure 2.11**c), and this state classification correlated with task performance both on subject and trial-by-trial level. Interestingly, the current state could also be decoded accurately from HPC activity, though unlike in OFC, task performance did not correlate with this decodability (Schuck and Niv, 2019), suggesting a unique role for OFC in representing decision-relevant task states. Only the hidden task-relevant information was decodable from OFC (**Figure 2.11**d), in accordance with the authors' hypothesis. Indeed, a searchlight analysis revealed that the only area in the brain from which all hidden task-relevant features were decodable was the medial OFC (**Figure 2.11**e). Similar results in rodents, suggesting mOFC is crucial for retrieving action outcomes specifically when task variables are unobservable, were obtained by (Bradfield et al., 2015).

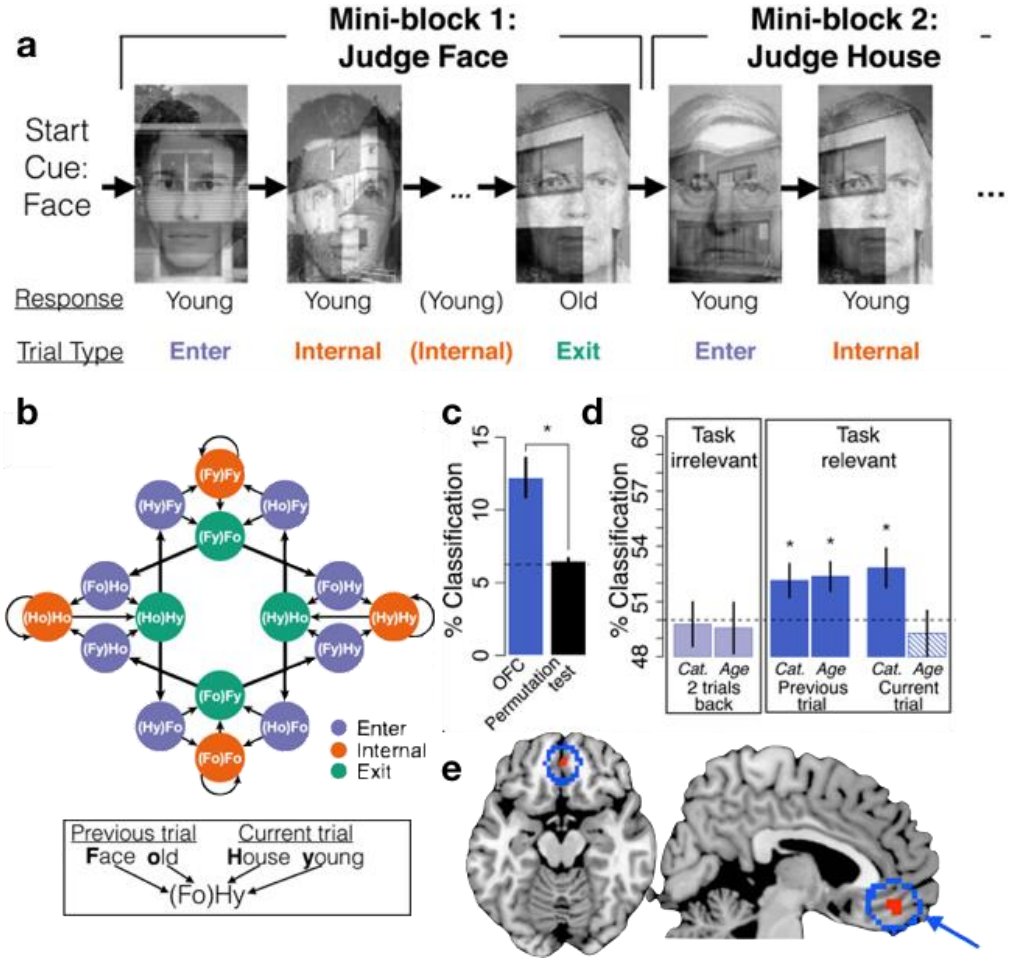


Figure 2.11 OFC codes the current (latent) state in an abstract task. a. Example of trial sequence in the task. Subjects began by judging the age (young versus old) of the cued category (face or house). In the following trials, they continued to judge the age of the same category until an age change occurred. On the trial following an age change, subjects had to switch to judging the other category, with the first trial after the change determining the new age of that category to which subsequent trials needed to be compared. These rules created an alternating mini-block structure of judging either the age of faces or houses. The first trial after a category switch was the trial in which the mini-block was entered (Enter trial), trials in which the category and age repeated are denoted Internal trials, and the trial in which the age changed was the end of the current mini-block (Exit trial). **b.** Task state space and possible transitions. **c.** Classification of state identity from the entire OFC. **d.** Only the three hidden state components contributed to accurate OFC state classification. **e.** The only brain regions to code for a conjunction of these components was the medial OFC. Copied from (Schuck et al., 2016).

Importantly, this account, or at least a weaker version of it suggesting the OFC compresses or separately represents states based on whether distinguishing them is important for behaviour (i.e. without emphasising unobservable components), is not at odds with the ubiquitous value representations in OFC. Beliefs regarding the task state and structure are critical to determining the value of the current trial, while the value of the current trial is a critical component of the underlying task state. Indeed, (Zhou et al., 2019) used an odour sequence task analogous to a spatial maze to show that though OFC representations were dominated by value signals, once these were regressed out, the specific location/state (both observable and latent/defined by memory) the animal was in was decodable with high fidelity. This coding of non-value dimensions of the task predicted behavioural mistakes, i.e. when the rat got confused regarding the current state.

While the jury might still be out for the strong formulation of this hypothesis (OFC is coding for the *current* state information, only if the state is *hidden*), it has certainly been useful and influential by emphasising the importance of OFC for state definitions. Knowledge of the state space is orthogonal to the MF/MB/MS division. It is needed for all types of RL learning algorithms. For example, the ability to learn causally in RL tasks (Jocham et al., 2016; Rudebeck et al., 2017; Walton et al., 2010) is reliant on correct knowledge of causal structure of the learning problem, which is a feature of the state space. Understanding the structure of the tasks (i.e., its state-space and transition function) is exactly what Harlow was describing as a learning set (Harlow, 1949). A learning set can be complex like "if A was not rewarded, it means that B is going to be rewarded" or as simple as "the event that just happened (reward) is most probably causally connected to the event that happened just before it (stimulus)" (Behrens et al., 2018a).

In the next chapter, I discuss one particularly important and well-studied structure: the structure of 2D physical space. Candidate representations for this state-space have been suggested – grid cells, originally found in the entorhinal cortex. I will discuss evidence for grid-like coding in frontal areas like vmPFC and (not conclusively) OFC, and intriguing similarities between coding properties of grid cells in physical space and OFC neurons in economic choices. In chapters 5 and 6, I show that the same regions where grid cells are found also code for the structure of an RL task (a variant of a 3-armed bandit or reversal learning task). These results bring us one step closer to an intimate understanding of the representations of task state-spaces in abstract RL tasks.

3 COGNITIVE MAPS OF RELATIONAL SPACES IN THE HIPPOCAMPAL FORMATION, VENTRAL PFC AND THE DEFAULT MODE NETWORK

Imagine you have just walked into a room you have never seen before. A large amount of prior knowledge is readily available for you to guide your behaviour: you know that the doors are the only comfortable exits from the room, or that behind the curtains there probably are windows. You know something about how rooms work: about the relations between entities in rooms. More basically, you know that to get to the distant corner of the room you will first need to go through the locations that are currently closest to you. You know something about the common structure of all two-dimensional spaces: a 2D topology induces specific constraints on the relations between locations. This relational internal model, termed "cognitive map" (Tolman 1948), and our remarkable ability to use it to guide behaviour in novel situations, is not unique to the spatial domain (Tolman 1948, Behrens 2019). Indeed, generalisation is a widespread phenomenon in human and animal behaviour (Tenenbaum 2011, Piaget 1926, Bartlett 1932), for example to support concept learning (Xu & Tenenabaum 2007), causal learning (Gopnik 2004) and reinforcement learning (Harlow).

In this chapter, I will first describe the exquisitely precise neural correlates of the spatial cognitive map found in the hippocampal formation, like place and grid cells. Next, I will present a growing body of evidence suggesting this type of coding is not restricted to the spatial domain and can be found in non-spatial tasks as well. In particular, evidence for grid-like coding has been found in a network similar to the default-mode network. I will then explain recent proposals that grid cells encode the relational statistical structure of 2D physical space, suggesting they are useful for spatial generalisation, as in the example above. Finally, I will discuss highly related work from the learning and memory literature, suggesting the default mode network is involved in the coding of knowledge structures called schemas.

This chapter is largely based on sections from a review I co-authored (Behrens et al., 2018a).

3.1 **A cognitive map of physical space**

3.1.1 The hippocampal zoo

Tolman suggested that when solving spatial tasks his rats relied on "complicated and autonomous intervening brain processes" (Tolman, 1948). The first evidence for such processes was discovered almost 30 years later in the hippocampus (O'Keefe and Nadel, 1978). The exquisite precision with which map-like representations are reflected in the firing of single neurons in the hippocampal-entorhinal system could not have been imagined even by Tolman.

The most celebrated of these neurons are active at particular places in the environment. Hippocampal place cells usually fire in one location in the environment (O'Keefe and Nadel, 1978). Medial entorhinal grid cells fire at multiple place locations on a triangular lattice (Hafting et al., 2005). The grid cell population can represent spatial vectors and distances between different spatial locations (Bush et al., 2015; Stemmler et al., 2015). Both place cells and grid cells are organised anatomically on the dorsal-ventral axis (posterior-anterior in primates) according to the size of their firing fields: dorsal (posterior) cells have small firing fields while ventral (anterior) have large firing fields. For grid cells, this progression is not continuous: grid cells separate into discrete modules with similar grid field size, grid spacing

and grid angle. However, different grid cells within the same module can have different phases (Stensola et al., 2012).

These famous cells are accompanied by a zoo (**Figure 3.1**)of less celebrated but equally remarkable cells that reveal the organisation of "knowledge" in the map (Grieves and Jeffery, 2017), such as cells that encode the current head direction of the animal (Taube et al., 1990), band cells (Krupic et al., 2012) and cells that encode the vectors from the animal to rewards (Gauthier and Tank, 2018), goals (Sarel et al., 2017), borders (Stensola et al., 2012), objects (Høydal et al., 2019). In social situations, some hippocampal cells encode the locations of other animals (Danjo et al., 2018; Omer et al., 2018).

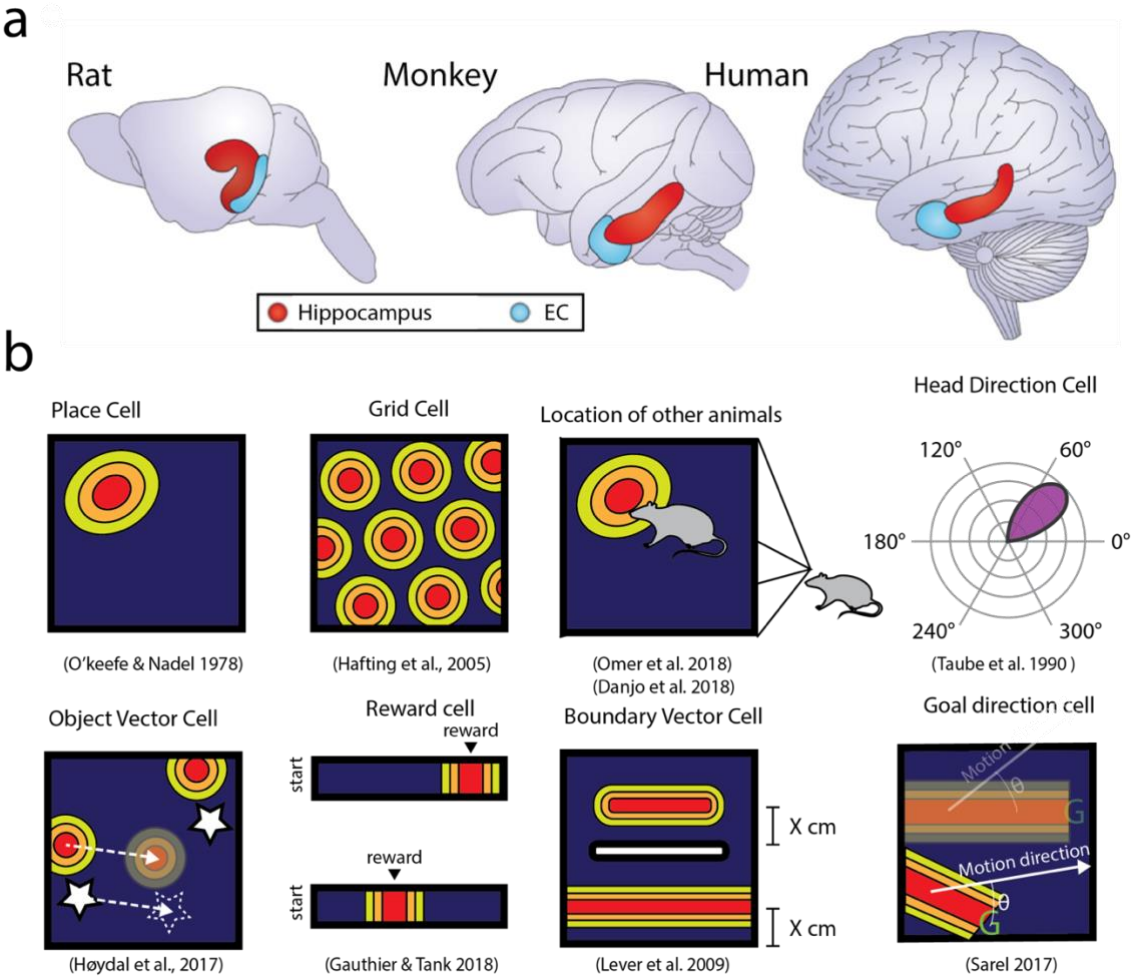


Figure 3.1 The hippocampal zoo. a. Anatomical location of the hippocampus and entorhinal cortex across species. Adapted with permission from (Strange et al., 2014). **b.** A variety of cells in the hippocampal formation represent different spatial variables. Place cells (O'Keefe and Nadel, 1978) are active when an animal is in a single (sometimes multiple) location. Grid cells (Hafting et al., 2005) are active when an animal is in one of multiple locations on a triangular lattice. "Social place cells" (Danjo et al., 2018; Omer et al., 2018) are active in one animal when it observes another animal is in a particular location. Head-direction cells (Taube et al., 1990) are active when an animal's head is facing a particular direction. Object-vector cells (Høydal et al., 2018) are active when an animal is in a particular direction and distance from any object. Reward cells (Gauthier and Tank, 2018) are active

when an animal is in the vicinity of reward. Boundary vector cells (Lever et al., 2009) are active at a given distance away from a boundary in a particular allocentric orientation. Goal direction cells (Sarel et al., 2017) are active when the goal of an animal is in a particular direction relative to its current movement direction. The green 'G' indicates the goal location. Non-spatial cognitive maps in hippocampus and ventral PFC. From (Behrens et al., 2018a)

3.1.2 The Neural underpinnings of spatial generalisation: remapping experiments

How does the spatial generalisation presented in the introduction to this chapter happen? There is a clear prediction for neural representation that might underpin it, and our precise knowledge about the nature of spatial neural representations renders this prediction testable: the neural code should stay the same for all different sensory instantiations of environments with the same structure. Here, "structure" is defined as the statistical structure of the relationships between elements in tasks. In free-foraging spatial task, these elements are simply locations in 2D (Gustafson and Daw, 2011) (We explore these ideas more deeply in chapter 4). For rats, there are statistical regularities that govern all examples of spatial free-foraging in a box. For humans, similar regularities govern all rooms (including more elaborated meanings of objects like windows and doors). Hence, a generalisable representation should code similarly for all 2D environments (at least for open-field, free-foraging tasks).

This prediction was tested in "remapping" experiments, where animals are moved between two different boxes (for example, with different wall colours and odours). In these experiments hippocampal place cells "globally" remap when the sensory environment changes (Bostock et al., 1991; Leutgeb et al., 2005). Neighbouring place cells in one environment are unlikely to be neighbours in the other[1]. This means the hippocampal code does *not* generalise across different examples of 2D space. By contrast, grid cells do not remap between environments, except for a rigid body transformation. Within a grid module, phase neighbours in one environment are also phase neighbours in the other (Fyhn et al., 2007)(**Figure 3.2**a) – the correlation structure in the population is preserved. The entorhinal

[1] When the sensory fratures of the environment change but not in a "substantial enough" manner, hippocampal placewcells maintain their place coding but display "rate remapping", whereby their firing rate ar their place field changes substantially between environments. In such cases, the grid cell population also maintains its spatial fields, without any rotation or translation (Fyhn et al., 2007).

representation of location therefore embeds the structural information about general relationships in 2D space that is common among all environments. Similarly, object vector cells in entorhinal cortex activate for any object present in the environment (Høydal et al., 2019), whereas similar cells in hippocampus activate for only a subset of objects (Deshmukh and Knierim, 2013)(**Figure 3.2**b).

Importantly, when the structure of the task changes from free foraging to a spatial task with a different structure, the grid code *does* change (Boccara et al., 2019; Butler et al., 2019). For example, when rats are trained to daily learn three new reward locations on a cheeseboard maze, the entorhinal (and hippocampal) code changes throughout the learning, with grid (and place) fields moving towards rewarded locations to represent the statistics of the current task, leading to long-lasting changes in the grid map. This change of the grid code occurs through "flickering" between the representations of the old and new task – as if a process of inference over the task is occurring (Boccara et al., 2019) (**Figure 3.2**c-d). This result is highly related to an earlier hippocampal "flickering" result from a sensory remapping experiment, where animals were "teleported" between two sensory instantiations of a room (Jezek et al., 2011). While the hippocampal code shows both sensory (Fyhn et al., 2007; Jezek et al., 2011) and task (Boccara et al., 2019) remapping, the entorhinal code showed only task remapping (Boccara et al., 2019). These results suggest grid cells (or EC cells in general) are useful for generalisation in space.

A similar phenomena to spatial remapping happens in OFC, but in a non-spatial space of food rewards: when monkeys are asked to make economic choices between different amounts of two juices, A and B, OFC neurons encode a rich variety of value and task-related variables (Xie and Padoa-Schioppa, 2016) but when these two juices are replaced with two different juices, C and D, the same neurons encode the exact same variables for the new juices (Xie and Padoa-Schioppa, 2016).

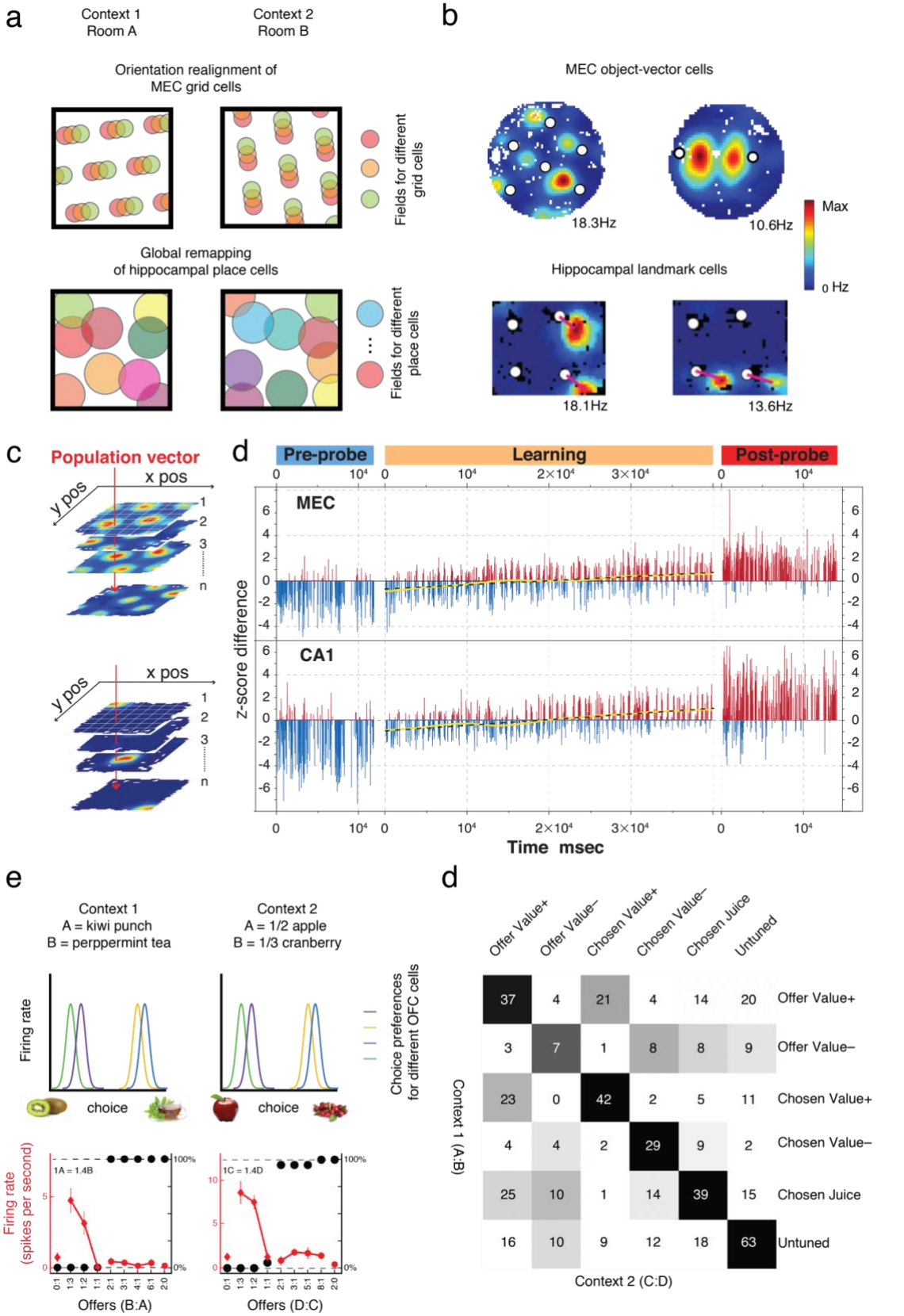


Figure 3.2 Generalisable neural codes: remapping experiments. a. In a spatial remapping experiment, animals are moved between two different environments ("contexts")_(Bostock et al., 1991; Leutgeb et al., 2005). Entorhinal grid cells maintain a constant spatial phase structure (top), in contrast with the global remapping of hippocampal place cells (bottom). **b.** MEC object vector cells

respond specifically when an animal is at a given direction and distance from any object, regardless of identity of the object (top) (Høydal et al., 2019). Cells with object-vector properties are also found in the hippocampus (bottom) (Deshmukh and Knierim, 2013). These cells, however, respond to only a subset of the objects (bottom). **c.** Illustration of the population vector calculation in different locations. This was done for both a pre-probe phase (free foraging, before the rewards are hidden in the new locations) and a post-probe phase (after learning of the new location. **d.** Difference between the instantaneous population vector correlations with the pre or post probe population vector. The representation changes towards the post-probe vector, but not smoothly: it flickers between the pre and post-probe representations. c. and d. adapted from (Boccara et al., 2019). **e.** When monkeys choose between pairs of different juices offered in different amounts, OFC neurons encode the same decision variables across contexts, such that two neurons 'supporting' the same decision for one pair of juices also support the same decision in different pairs of juices (top). Firing rate of an example OFC neuron as a function of the quantity ratio of the two juices (bottom). Black symbols represent the percentage of trials in which the monkey chose the less preferred juice. Red symbols represent the average firing rate. **f.** Population analysis demonstrating OFC neurons generalise over different contexts. Each task-related OFC neuron was classified according to its strongest encoded decision variable (offer value, chosen value, chosen juice) and the sign of this encoding. This was done in two separate contexts (A:B choice / C:D choice shown in rows/columns). Numbers are cell counts for encoding a pair of variables in the two contexts. Statistical significance was assessed via a null distribution (greyscale colours are P-values). The strong diagonal indicates neurons tend to encode the same variable with the same sign in both contexts. e. and f. adapted, with permission, from (Xie and Padoa-Schioppa, 2016). Panels a,b,e,f adapted from (Behrens et al., 2018b).

3.1.3 The hippocampal spatial code: more than an allocentric map

In the last section we saw how the hippocampal population code changes across either different sensory examples of the same task, or across different tasks. But what about hippocampal single cell representations? Are they also sensitive to the task, or do they solely represent the animal's location in an allocentric map (or as the Noble prize committee put it – "the brain's GPS system")?

(Wood et al., 2000) developed a continuous alternation task in a modified T-maze, such that they traversed the central arm of the maze on each trial and then alternated between left or right turns (**Figure 3.3**a). Two-thirds of hippocampal place cells with firing fields in the common arm fired in different ways, depending on the trial types (**Figure 3.3**b). (Ferbinteanu and Shapiro, 2003) showed that these neurons usually coded for the last choice that was made rather than the choice that is going to be made. Similarly, when a rat running on a circular track receives reward every 4 loops, some place cells will fire at a location only at specific loop numbers (e.g. every time the rat reaches the location during the second round since the last reward was received), as that is the variable that is relevant for the task (Sun et al., 2020).

Another example of a hippocampal coding that does not fit with the original ideas of place cells as a completely allocentric mapping of the environment is the experience-

dependent asymmetric shape of place fields (Mehta et al., 2000). When rats are trained to prefer a running direction on a narrow linear track, place fields skew backwards - away from the preferred running direction (**Figure 3.3**c). An elegant explanation for these results was recently given by (Stachenfeld et al., 2017), suggesting the hippocampus codes a predictive, rather than purely spatial map. (Stachenfeld et al., 2017) harnessed RL theory (see a thorough explanation in chapter 4) to show how the backwards skewing of place fields, together with many other hippocampal results, can be obtained if the momentary population vector of place cells is thought to be coding a single row of the successor representation $M(s, s')$ (section 1.3.6.1). The row corresponds to the current state s the animal is occupying in the task space. Each neuron in the vector (indexes running along the column of the SR) represents a possible future state (for example a spatial location). The firing rate of a single neuron encoding a state s' is proportional to the discounted expected number of times s' will be visited, assuming the current policy, and given the current state s. For example, the dependence on the task state-space definition can explain how splitter cells can arise: even though the central arm has the same perceptual appearance in both left and right-turn trials, the task states are different. I further discuss the hippocampal SR theory in section 3.2.3.1.

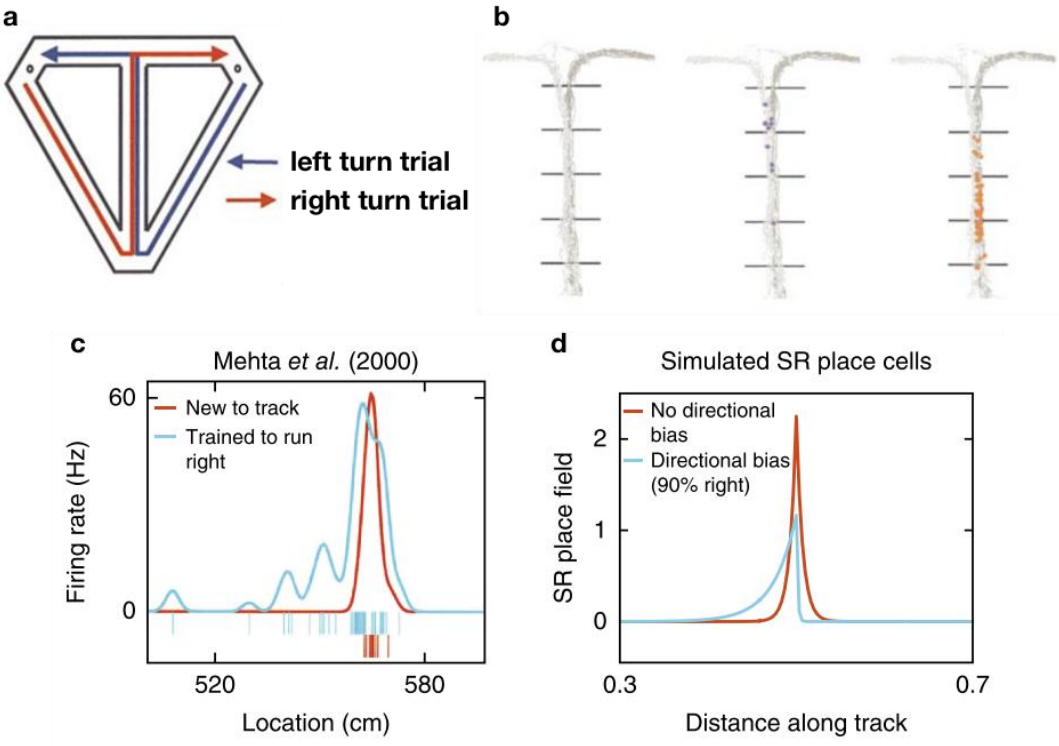


Figure 3.3 Behaviourally-dependent changes in place fields. a-b: Splitter cells. **a.** maze structure. **b.** Dependency of firing of an example hippocampal cell on trial type. The paths taken by the animals on the central stem are plotted in the left panel (light grey, left-turn trial; dark grey, right-turn trial). In the middle and right panels, the location of the rat when individual spikes occurred is indicated

separately for left-turn trials (purple dots) and right-turn trials (orange dots). Adapted from (Wood et al., 2000). **c.** Backwards skewing of place fields. Adapted from (Mehta et al., 2000). **d.** Simulated successor representation on the task from (Mehta et al., 2000) produce the same backwards skewing. c. and d. adapted from (Stachenfeld et al., 2017).

Another dimension of complexity of the HPC-EC spatial code is its functional gradient along the dorsal-ventral (posterior-anterior in humans) axis of place fields size. While in the dorsal hippocampus place fields are abundant and tend have small place fields, the anterior hippocampus has fewer place fields which tend to have large place fields (Jung et al., 1994). This diversity in representational resolution is mirrored in grid fields of the entorhinal cortex, with the same dorsal-ventral functional gradient. Grid cells, however, are organised in discrete modules of grid field size, unlike hippocampal place fields which seems to vary continuously (Hafting et al., 2005).

3.1.3.1 Replay and preplay in physical space

Although place and grid cells predominantly code the animal's current location, they also spontaneously replicate temporally-compressed previous trajectories (Diba and Buzsáki, 2007; Foster and Wilson, 2006; Ólafsdóttir et al., 2016; Wilson and McNaughton, 1994) and explores possible future and even new trajectories (Dragoi and Tonegawa, 2011; Freyja Ólafsdóttir et al., 2015; Gupta et al., 2010) that have not been experienced, both during sleep and (awake) rest (Karlsson and Frank, 2009). These phenomena are called *replay* and *preplay*, respectively (Foster, 2017).

Replay is critical for behaviour (Jadhav et al., 2012). A prominent suggestion is that replay is crucial for memory consolidation, as it enables co-activation on synaptic time scales of cells, representing events happening orders of magnitude more slowly. Theoretical proposals have suggested that the hippocampus acts as a "teacher", replaying experiences for memory consolidation in cortex (McClelland et al., 1995). Indeed, during hippocampal "sharp-wave ripples" (high frequency activity, when most replay and preplay activity is found, though see (Johnson and Redish, 2007)), most of the cerebral cortex is selectively activated, while most diencephalic, midbrain and brainstem regions are strongly and consistently inhibited (Logothetis et al., 2012). Interestingly, after obtaining reward, the replay sequence reverses (and plays from the goal to the start location) – suggesting a mechanism for credit assignment (Ambrose et al., 2016).

The ability to explore future (and even unexperienced) parts of the task during pre-play has been suggested as important for building and sampling internal models of the

environment, and as a possible neural substrate for goal-directed behaviour or model-based planning (section 1.3.4 and section 2.4).

3.2 Non-spatial cognitive maps in the medial temporal lobe and ventral PFC

3.2.1 Relational memory and the hippocampal formation

A seemingly separate line of work on hippocampal formation function began with the study of amnesia following hippocampal damage. Perhaps the most famous case was of patient H.M., who suffered from anterograde and short-term retrograde amnesia following the removal of his bilateral hippocampi and anterior parahippocampal gyrus (including the entorhinal cortex)[1] for treatment of drug-resistant epileptic seizures (Scoville and Milner, 1957). H.M.'s perceptual and intellectual functions seemed perfectly spared, suggesting (contrary to the belief at the time) that memory is a distinct function which is not widely distributed in cortex. In addition, some of H.M.'s memory functions were spared: his performance in visuomotor tasks like mirror reading improved with practice, even though he did not consciously remember he ever practiced. This led to a suggestion of a distinction between two types of memory: procedural memory and declarative memory, where only the latter is hippocampal formation dependent (Cohen and Squire, 1980).

These early suggestions focused on the human amnesia literature, using the literal meaning of the word "declarative" – the ability to remember ("declare") facts and events (Cohen and Squire, 1980). These ideas were further developed in the seminal synthesis by (Cohen and Eichenbaum, 1993), who gave a broader, yet more precise definition of declarative memory, trying to encompass a wide range of hippocampal results across species and domains. They suggested the hippocampal formation provides a general relational processing mechanism which enables relational representations that bind the elements of experiences and links memories through their common elements (Cohen and Eichenbaum, 1993; Eichenbaum, 1999). Associations between inputs (or subsets of inputs), and between existing relational

[1] Later reports of a different patient confirmed the same symptoms could be found following a more restricted lesion only including the CA1 hippocampal field (Zola-Morgan et al., 1986).

memories and inputs, are rapidly formed in the hippocampus. This renders the hippocampus critical for the generation (e.g. imagination: (Hassabis et al., 2007)), recombination (e.g. transitive inference or pair-associated paradigms: (Bunsey and Elchenbaum, 1996; Dusek and Eichenbaum, 1997; Preston et al., 2004; Zeithamova et al., 2012)) and flexible use (e.g. abstraction of commonalities between experiences and generalisation: (Koster et al., 2018; Kumaran et al., 2009; Kumaran and McClelland, 2012)) of information (Eichenbaum and Cohen, 2014).

The relational binding of different events in different episodes through their common elements composes a "memory space" – similar to physical space where different locations, which can be part of different routes, are composed into a cognitive map (Eichenbaum et al., 1999) (**Figure 3.4**.). In fact, physical space can be viewed as a particular example of a memory space, in tasks where the emphasis is purely spatial. Indeed, in a task where rats learn multiple context-dependent object-reward-location associations, the hippocampal population code can be characterised as a hierarchy of relations between task dimensions – with other dimensions such as the context, item identity and valence being coded besides the spatial location. Importantly, after initial learning of one set of associations, subjects were much faster in assimilating new associations into the relational associative network already established (Mckenzie et al., 2014).

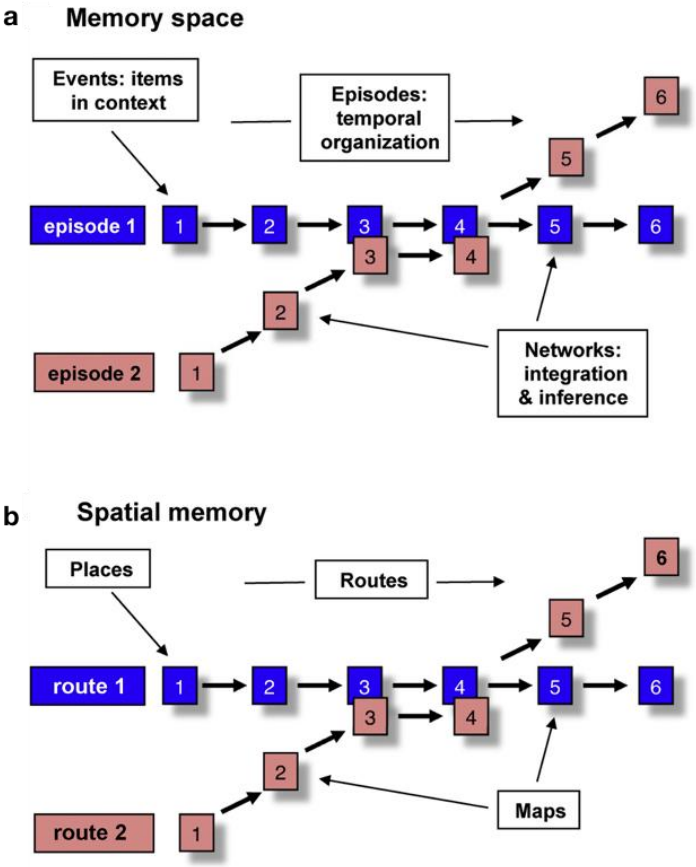


Figure 3.4. Memory spaces and their connection to physical space. a. In the general case, items (object, behaviours, and events) are positioned within the context in which they occurred to make up events. Episodes are composed as a temporal organization of events. Relational networks are composed as links between events and episodes, supporting the ability to make inferences between related events that were not experienced together. **b.** Spatial memory as an example of a memory space. The events are now comprised only of places, the episodes from routes, from (Eichenbaum and Cohen, 2014)

3.2.2 Place and grid-like non-spatial representations

Hence, the hippocampus is important for different types of relational reasoning – not just physical space. But what does that tell us about the representations of single cells in the hippocampus? Can we use our understanding of hippocampal formation representations in spatial tasks to examine representations in other relational tasks?

The term "map" in 2D physical space entails obvious and specific predictions for representations. The fact that physical space is a very important domain for rodents has probably also contributed to the interpretation and early discovery of the physical cognitive map. Human place cells have been reported (Ekstrom et al., 2003), but it is notably more difficult to find place cells in human and non-human primates' hippocampus (Rolls, 1999).

This suggests that the primate cognitive map, while encoding spatial features, is less dominated by spatial representations than in rodents.

Even the rodent's hippocampus primarily encodes non-spatial features in non-spatial tasks (Wood et al., 1999) or partially spatial tasks (Mckenzie et al., 2014). However, the possible predictions for neural representations in such tasks don't allow the precision of prediction possible in the spatial domain. In contrast, continuous 1D or 2D non-spatial domains have very similar features to physical space, and hence make similarly precise predictions. Indeed, a growing number of studies show activity in the same regions where place and grid cells have been found, measured either directly or with fMRI, which show analogous coding of non-spatial information.

One obvious 1D dimension which is highly relevant for all animals is time. Indeed, when time, rather than space, is important for the task, some hippocampal place cells exhibit "place cell-like" coding of time. For example, when animals need to run on a treadmill for a specific amount of time, a subset of hippocampal cells tile the time interval, with each cell coding for a specific time during the interval (Kraus et al., 2013). Meanwhile, entorhinal grid cells showed the same pattern but with multiple peaks (Kraus et al., 2015) – like grid cell spatially-tuned firing when an animal is moving along a linear track (Yoon et al., 2016). Importantly, the location of the treadmill was fixed, and the speed of the treadmill varied, such that the distance travelled (and hence also the path integration calculation) varied between trials (for a review, see (Eichenbaum, 2014)). These hippocampal time representations might be critical for organising memory in time, in congruence with abundant lesion literature suggesting the hippocampus is critical for memory of the order of sequential events, even when it was not needed for remembering the specific events themselves (Eichenbaum, 2013).

Similar results can be obtained for arbitrary sensory dimensions, addressing any concern about the "uniqueness" of space and time. (Aronov et al., 2017a) trained rodents to hold a lever whilst a tone increases in frequency and then release it at a target frequency to get a reward - a non-spatial, 1D operant continuous task. Hippocampal cells exhibit place-like firing fields, but for different frequencies rather than places (**Figure 3.5**d bottom); again, entorhinal cells (including a third of spatial grid cells) showed multiple distinct frequency fields (**Figure 3.5**e bottom), much like those of grid cells when an animal is running in a linear track (Yoon et al., 2016) (**Figure 3.5**d-e top), though a direct correspondence was not assessed.

Grid-like activity can be recorded in humans while they navigate a virtual reality environment, either directly from the activity of cells in the entorhinal cortex (recorded for medical reasons) (Jacobs et al., 2013) or indirectly with fMRI. While fMRI does not offer direct evidence of the neuronal firing patterns, the hexagonal symmetry of the grid pattern has a striking signature in the fMRI signal. The BOLD signal exhibits a 6-fold oscillation as a function of running direction in the VR environment (Doeller et al., 2010) (**Figure 3.5**a). This signature is sometimes referred to as a "hexa-directional signal".

Critically, this pattern was also observed when subjects performed an operant non-spatial task that has the same statistical structure as navigation in physical space (2D continuous organisation (Constantinescu et al., 2016). Subjects watched cartoons of birds as they morphed in two dimensions – the lengths of the legs and the necks. The task was to predict when the appearance of the birds matches one of several learned target birds. The instantaneous change in the appearance of the bird defines a vector in 2D conceptual "bird-space". The methods described above for testing of grid-like coding were then applied, i.e. looking for 6-fold oscillations as a function of this vector (instead of the running direction in the VR). This analysis showed an effect in the entorhinal cortex (EC) and medial prefrontal cortex (mPFC) (**Figure 3.5**c), both areas where grid cells where recorded directly in humans using VR (Jacobs et al., 2013). The network showing this grid-like coding was actually wider (though without the rigorous statistical tests used for vmPFC and EC), and was remarkably similar to the rest of the network reported by (Doeller et al., 2010) (**Figure 3.5**b) for spatial grid-like coding, including posterior cingulate cortex (PCC), retrosplenial cortex (RSC), temporal-parietal junction (TPJ) and posterior parietal cortex (PPC). Similar signals have been observed in a 2D continuous task where the dimensions are given by two odours, mixed together to produce a "location" in 2D odour space (Bao et al., 2019). This network is strikingly similar to the default mode network (DMN) – as I discuss further in section 3.3.1.

By contrast, in hippocampus, cells show selective activity to specific abstract stimuli, such as (anything related to) Jennifer Aniston (Quiroga et al., 2005) – a specific "location" in the conceptual space, analogous to place cells in spatial domains. This is also true also in a decision-making task where monkeys need to choose between 3 options with varying reward probabilities. Here, the task space is 3D and is defined by the value of each of the options. Hippocampal cells fire only in a particular location in this space (Knudsen and Wallis, unpublished observations).

The hexa-directional signal was also found in 2D discrete, rather than continuous spaces. In humans navigating a 2D discrete "social" space where the dimensions are "competence" and "popularity" of individuals. Though the dimensions were always learned separately, trajectories through this 2D space elicited a hexa-directional signal in the same network as in (continuous) physical and bird spaces (Park and Boorman, unpublished observations). This fMRI effect was again observed in vmPFC in non-human primates making choices in a 2D value space defined by the probability and magnitude of reward (Bongioanni et al., 2020). In addition, grid-like hexa-directional signals have been shown to encode gaze location on a 2D image in both humans (Julian et al., 2018; Nau et al., 2018), a finding that was corroborated by hippocampal recordings in monkeys (Killian and Buffalo, 2018).

Another spatial property of the HPC-EC system that seems to transfer to non-spatial spaces is the functional gradient in firing field sizes. While dorsal HPC neurons rapidly associate event identities with spatial locations, anterior HPC neurons accumulate evidence over longer time-scales and generalise across related events – mirroring the high to low spatial resolution of place cells along the HPC long axis (Komorowski et al., 2013). This finding was replicated in humans along the homologous posterior-anterior axis (Collin et al., 2015). These results are particularly interesting for generalisation, as the anterior HPC is the part that is connected to frontal cortex, and in particular vmPFC.

Hence, there is evidence that place and grid activity patterns are neither unique to physical space nor, in humans at least, unique to the hippocampal formation. Instead, in physical space they may reflect the 2D topology that is inherent to space. These results suggest that the role of place and grid cells in spatial cognition is a specific instance of more general coding mechanisms realized in the hippocampal formation and the default mode network.

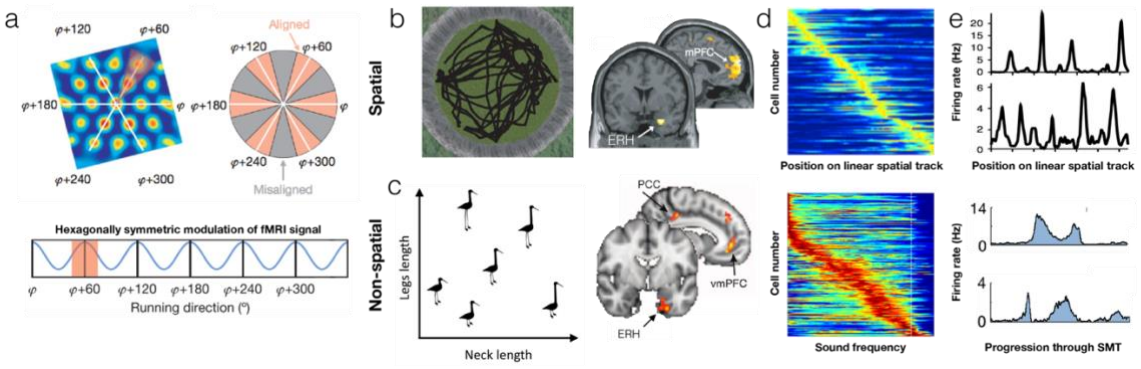


Figure 3.5 Generalised spatial representations. a. Logic behind measuring grid cells with fMRI. Trajectories through a 2-dimensional space can be either aligned or misaligned with grid axes (white lines denote grid axes). Greater signal for trajectories aligned vs misaligned with the grid axes results

in a hexagonally symmetric sinusoidal modulation of the BOLD signal as a function of running direction. Adapted, with permission, from (Doeller et al., 2010). **b.** This "grid-like" hexa-directional signal is observed in entorhinal cortex (EC) and mPFC (right) when participants navigate through spatial virtual reality (Doeller et al., 2010). Left is an aerial view of the task map. **c.** The same signal is observed when participants navigate through an abstract conceptual space defined by the neck length and leg length of "stretchy birds", suggesting grid cell code exists in non-spatial 2D tasks (Constantinescu et al., 2016). **d.** Cells in the hippocampus fire at specific sound frequencies in a non-spatial, sound manipulation (SMT) task (bottom) in a manner analogous to the representation of spatial locations in place cells on a linear track (top). In both panels, each row presents the normalised activity of one hippocampal cell, as a function of either distance along a linear spatial track or sound frequency. Cells are ordered according to the position of the firing field. Top adapted from (Miao et al., 2015) and bottom from (Aronov et al., 2017b). **e.** Grid-like coding in EC of a linear spatial track (top) and progression through the sound manipulation task (bottom). Top adapted from (Yoon et al., 2016) and bottom from (Aronov et al., 2017b). Figure from (Behrens et al., 2018a).

3.2.3 Other features of hippocampal non-spatial tasks maps

There are other theories and observations of hippocampal representations that were originally applied in spatial tasks and have proved useful in abstract and discrete relational tasks as well.

3.2.3.1 Maps of discrete abstract state-spaces and the successor representation

The suggestion by (Stachenfeld et al., 2017) that hippocampal activity could be explained as coding for the SR was not limited to explaining the firing fields of spatial place cells. (Schapiro et al., 2013) presented subjects with sequences of stimuli drawn from a graph with a community structure – i.e. with distinct clusters of interconnected nodes, and sparse connections between the clusters. The hippocampal BOLD signal was more similar within community than across communities – a property that is predicted by the SR embedding of the task (i.e. a row of the SR, assuming each stimuli/node in the graph is a state). This happens because the agent is more likely to visit sooner other nodes in the current community than nodes in other communities. Similarly, the SR predicts the BOLD adaptation signal in the hippocampal formation for a graph, even if the subjects are not consciously aware the task has any structure (Garvert et al., 2017).

3.2.3.2 Replay and preplay in abstract tasks

Recently, advances in analyses methods, combined with clever experimental designs, have provided evidence that replay and preplay are phenomena that also occur in the human brain. Moreover, they occur in tasks with abstract state spaces, where the sequence that is being played is a sequence of representations of task relevant states (instead of representations of locations). We explore the idea of treating locations as states in an MDP further in chapter 4.

(Schuck and Niv, 2019) recorded fMRI from HPC and OFC while subjects learned and performed the abstract state-space task from section 2.6.4 ((Schuck et al., 2016)), and also during a rest period after the completion of the task. They first show that despite the sluggish nature of fMRI, sequentially replayed events on the (human replay) order of a few hundred milliseconds can be decoded from simulation data. Next, they applied the classifier trained on HPC data during the task to each volume (TR) of the resting data, resulting in a sequence of decoded states. They then show that these decoded rest sequences were related to the distances between states in the task, such that consecutively decoded states were nearby in task space. Through a battery of control analyses and simulations, the authors show that this "sequenceness" of decoded states could in theory result from sub-second replay events, and that it was specific to the rest period following completion of the task (compared to a rest period before exposure to the task or to shuffled data) suggesting it was not a by-product of the sequential nature of the task. Taken together, these results suggest that hippocampal replay occurs in humans, and encompasses abstract tasks.

A similar conclusion, with some very interesting extensions, was reached by (Liu et al., 2019) who also studied human brain representations during rest by training classifiers on data collected while subjects were completing an abstract task similar to transitive inference (unjumble a jumbled sequence of stimuli pairs) – this time using MEG (magnetoencephalography) rather than fMRI. They too used a measure of "sequenceness" to show that order of sequences of decoded states (here decoded from the entire brain, due to MEG's poor spatial resolution) followed a task-relevant order rather than the order of sensory experiences. These sequences shared many features of rodent sharp-wave ripples replay: they were time-compressed, localised (as far as localisation goes in MEG) to the hippocampus, and occurring specifically at times with a strong power increase of the sharp-wave ripples frequency (100-180Hz). In addition, as in rodents (Ambrose et al., 2016), while the sequences usually played in the forward direction, after the last state was rewarded the sequence was played backwards – a mechanism for credit assignment.

However, they did not just look for evidence of replay, but also for mechanistic insight into the role of replay in the organisation of experience according to abstract knowledge. When well-trained subjects were confronted with a new set of stimuli pairs to unjumble, the replay of the position of the picture (trained on different stimuli, according to their position in the sequence) consistently led the replay of picture identity. This suggests that the abstract learned rule was critical for placing the correct item for the current place in a sequence.

Hence, replay is one mechanism by which abstract structural knowledge can efficiently and usefully organise novel experience.

3.3 **Hippocampal formation and vPFC: similarities, differences and interactions**

The OFC/vmPFC is the prefrontal region that is most extensively interconnected with the hippocampal formation (Cavada et al., 2000). The ardent reader will notice many similarities between proposed functions of structures in the hippocampal formation and the OFC in the flexible encoding of complex features of the environment in a way that preserves relational information (Behrens et al., 2018b; Wikenheiser and Schoenbaum, 2016). Both the EC and the central OFC maintain their covariance structure when the sensory stimuli in tasks change but the structure of the task is the same (section 3.1.2 and **Figure 3.2**). Structural knowledge of the relationships between elements in the task is encoded both in EC (as is further explored in chapter 5) and OFC (as is evident from the learning set effects discussed in sections 2.6.2 and 2.6.4). EC and HPC both encode location in physical space, with OFC having at least some (though sometimes weak) representation of spatial location encoded. Both OFC and HPC (and probably EC, though this has not been reported) encode the current location in the abstract task described in section 3.2.3.2 ((Schuck et al., 2016; Schuck and Niv, 2019)). A related effect was present in HPC, EC and mOFC which all show multivariate activity patterns that are linearly related to Euclidian distance in an abstract (social) discrete 2D task, even when the two dimensions are learnt separately (Park et al., 2019).

However, (Schuck and Niv, 2019) have also found interesting differences between the two representations. While the coherence of OFC current-state (online) representation correlated with the performance of subjects in the task, there was no such correlation for HPC current-state representation. In addition, offline HPC replay (as indexed by the "sequenceness" during rest) correlated with the strength of the OFC representation, suggesting HPC replay helps to form or solidify OFC's behaviourally-relevant online representations. Relatedly, without a major hippocampus output (ventral subiculum), rats were slower to adapt to changes in reward contingencies, and OFC representations of expected outcomes, outcome features and response information were degraded (Wikenheiser et al., 2017). Importantly, the interaction between HPC and OFC is

bidirectional, for example with (lateral) OFC learning signals like prediction errors (with the subject's goals as the frame of reference) influencing hippocampal maps of specific stimulus-outcome associations, as indexed with BOLD repetition suppression (Boorman et al., 2016).

To summarise this suggested view: online, decision-related representations in OFC (which are more grounded in the current needs of the animal), are supported by (but also exert influence on) ongoing, promiscuous associative relational processing in the hippocampal formation. Evidence for this view can also be found in sensory preconditioning experiments. Deficits in sensory pre-conditioning, which are usually associated with lesions in hippocampal formation structures, are also caused by OFC lesions, though with an important double dissociation: OFC was found to be important for inference at the test stage, while intact hippocampal formation was important at the encoding (pre-conditioning) stage (reviewed in (Wikenheiser and Schoenbaum, 2016)). This account can also explain why in a task where the reward associations of events depend on the context in which they occurred, OFC representations are dominated by value signals (the most important feature of the task for behaviour), but are also sensitive to other behaviourally important variables like the appearance of an object in a specific context (Farovik et al., 2015), whereas in the same task, HPC representations are governed by more general associative information like context and position which can help build the associative map, and only show a weaker coding of the value of different states (Mckenzie et al., 2014).

3.3.1 Schemas

Schemas are defined in the learning and memory literature as superordinate knowledge structures that reflect abstracted commonalities across multiple experiences, exerting powerful influences over how events are perceived, interpreted, and remembered (Gilboa and Marlatte, 2017). While this definition sounds remarkably relevant to the topics of generalisation, learning sets and structural representations we have discussed so far, connections between these lines of work are surprisingly sparse.

In many of the goal-directed or relational tasks discussed above (like spatial navigation, imagination, inferential reasoning and autobiographical memory (Barron et al., 2013; Bowman and Zeithamova, 2018; Hassabis et al., 2007; Zeithamova et al., 2012)) hippocampal activity is accompanied by strong vmPFC activity. These two regions are part of the functionally connected default mode network (DMN), which also includes the posterior cingulate cortex (PCC) and the angular gyrus as well as other structures in the hippocampal

formation, retrosplenial cortex (RSC), tempro-parietal junction (TPJ), the temporal pole and lateral temporal cortex. Work on schemas in both rodents and humans has highlighted this network and in particular the hippocampus and the vmPFC/prelimbic cortex.

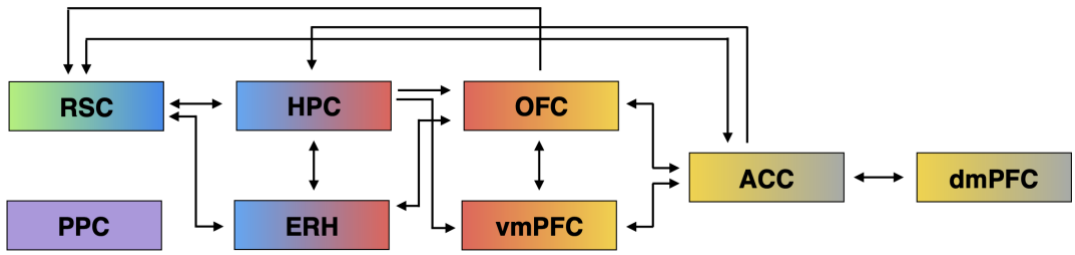


Figure 3.6 Anatomical connectivity of the DMN. Regions with similar colour gradient also have similar anatomical gradients. Note the OFC is not usually considered part of the DMN, except for its medial parts (BA14). Taken from Alexandra Constantinescu's DPhil thesis in Oxford. RSC: retrosplenial cortex; HPC: hippocampus; OFC: orbitofrontal cortex; PPC: posterior parietal cortex; ERH: entorhinal cortex; vmPFC: ventromedial PFC; ACC: anterior cingulate cortex; dmPFC: dorsomedial PFC.

In a classic set of experiments (Tse et al., 2011, 2007), rats were trained in a spatial 7x7 grid to memorise the locations of 6 different flavoured rewards hidden in sand wells, thus forming 6 flavour-place associations (PAs). When cued with a specific flavour at the start boxes, they need to go to the correct well. Probe tests were conducted in extinction: animals were given a flavour cue, but there was no actual reward in the well. The time the animals kept digging for the reward serves as an index of the strength of the PA. Expectedly, animals who had HPC lesions, could not form PAs. However trained animals who were exposed to just a single trial of two new PAs just before HPC lesion did learn and retained this knowledge post-surgery. This fast learning was enabled because of the existence of a schema (the other PAs) to assimilate the new PAs into. Indeed, when animals were trained in a similar way (same locations and flavours) but with a constantly changing mapping between specific rewards and locations (i.e. a schema of PAs could not be formed), they did not show the fast learning of new PAs.

A follow-up study (Tse et al., 2011) used this paradigm to investigate the role of HPC and mPFC (as well other control regions) in the process of assimilating information to and retrieving information from schemas. The hippocampal-dependent learning of new PAs was associated with a striking up-regulation of immediate early genes (marker for activity) in the prelimbic region of mPFC, specifically for animals that had to assimilate the new PAs into an established schema. Silencing this area resulted in deficits in new learning (supposedly because new PAs could not be assimilated into an existing schema) and the recall of

information. (Sommer, 2017) replicated these results using the same paradigm with fMRI in humans, showing stronger vmPFC (and RSC) activity when encoding new PAs that could be assimilated into an existing schema than PAs that could not. In chapter 6, I will investigate a somewhat similar idea, though in a very different context – the updating of the internal model of an RL task through prediction errors – a process that in many ways is the RL analogy to assimilation of new information into an existing schema.

Most of the studies of schematic cognition (including, to some extent, the one presented in chapters 5 and 6) have been conducted in highly controlled and artificial environments. However, we use schematic knowledge routinely in our everyday life. Imagine, for example, you walk into a new airport in a foreign country you have never seen before. Even though all the sensory stimuli are novel, you know what the actions are that this experience should include: entering the airport, going through the security checkpoint, walking to and waiting at the gate, and boarding the airplane. You can use your "airport schema" to guide your behaviour. (Baldassano et al., 2018) used an experimental design that is very similar to the remapping paradigm (discussed in section 3.1.2 and that I use in chapters 5 and 6), to investigate representations of such naturalistic schemas. They recorded fMRI while subjects where either watching movies or listening to stories depicting airport or restaurant experiences. This resulted in a factorial design of modality (movies/audio) vs experience/schema (airport/restaurant). They then used a representational similarity analysis (RSA, very similar to the one I use in chapters 5 and 6), together with fancy data analysis techniques that project neural responses from different subjects to a common low-dimensional space that aligns corresponding time points from the same stories) to test where in the brain there is a higher correlation between activity patterns of the same schema vs different schemas. They found such patterns across the entire default mode network, and in particular mPFC (**Figure 3.7**). This result held even when comparing only same-schema stories from different modalities (only diagonally crossed elements in the RDM in **Figure 3.7**a with significant brain regions shown by the diagonally crossed areas in **Figure 3.7**b). The authors could even predict the script (airport/restaurant) with high accuracy from DMN areas activity.

These results, together with those presented earlier and my own results that will be discussed later in this thesis, suggest that the same brain regions are involved in the representation of schematic and structural knowledge across a wide range of domains, including physical space, RL, and naturalistic environments.

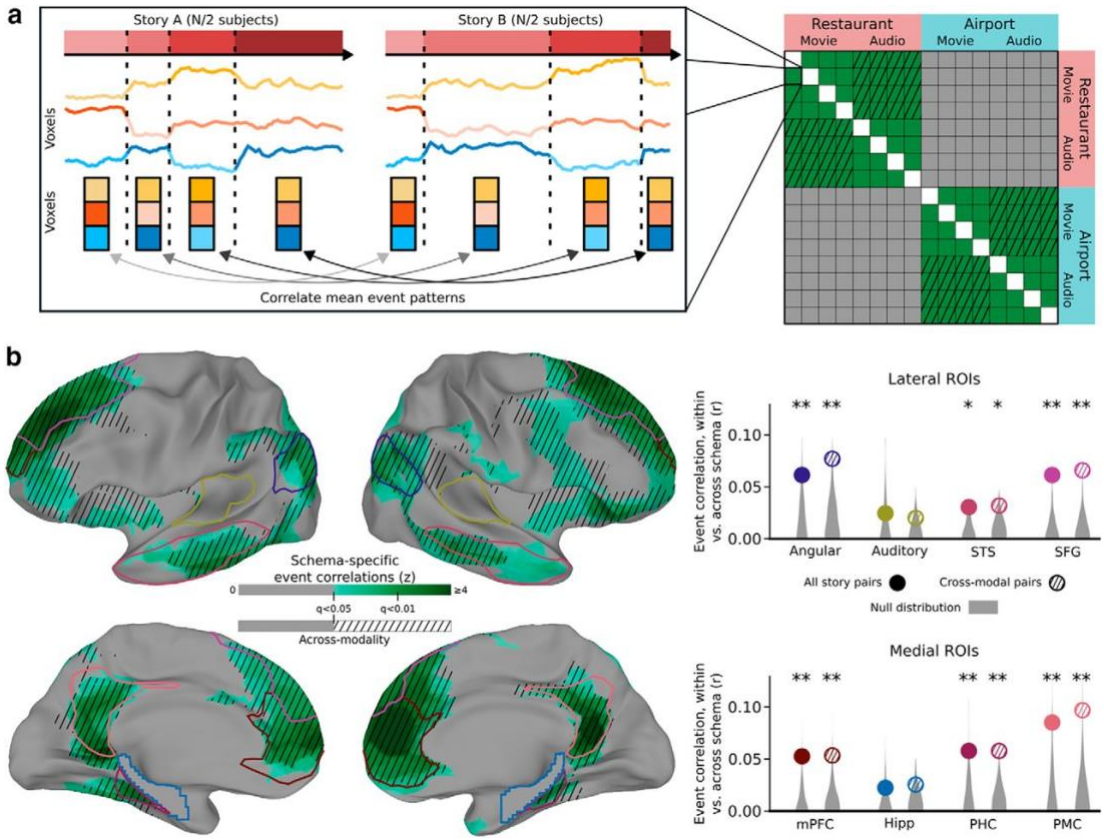


Figure 3.7 DMN represents stories with the same schema similarly. a. For each pair of the 16 stories, the mean pattern correlation (in a ROI) between the 4 corresponding subevents (different parts of the story). These correlations were then binned depending on whether the two stories were drawn from the same schema (green elements) or different schemas (grey elements). Comparisons which are strictly across-modalities are highlighted with stripy diagonal lines. **b.** Throughout the default mode network, stories from the same schema showed significantly higher event similarity compared with stories from different schemas. This result was not also true when only considering across-modality stories (audiovisual movie and audio narrative). STS: superior temporal sulcus. SFG: superior frontal gyrus. PHC: parahippocampal gyrus. PMC: Posterior medial cortex (i.e. PCC, RSC and the pre-cuneus cortex. From (Baldassano et al., 2018).

3.4 **Thesis aims**

In the introduction chapters I have reviewed two subfields of neuroscience/psychology: reinforcement learning and cognitive maps. In the experimental work in the rest of this thesis, I aim to use tools and ideas from one field to answer open questions in the other. In chapter 4, I use ideas from the RL world (namely MDPs and state-spaces) to examine the computational properties of spatially tuned neurons, and to suggest a novel efficient navigation algorithm that is useful in both spatial and non-spatial environments. In chapters 5,6 and 7 I use concepts and analysis techniques from the cognitive maps world (namely, the examination of the covariance of neural populations in remapping experiments) to investigate how abstract RL tasks can be represented and generalised.

C

Part 2: Experimental Work

4 PHYSICAL SPACE AS A STATE-SPACE WITH GRID-LIKE EIGENSPECTRUM, AND BENEFITS FOR INTUITIVE PLANNING

In 1948, Edward Tolman coined the term "cognitive map" to describe the encoding of the relationships between entities in the environment that supports flexible navigation. While Tolman's experiments were all spatial, he suggested that humans organise their general knowledge using similar relational cognitive maps. Place and grid cells provide evidence for such a map in physical space, and support flexible and efficient mechanisms of navigation in a spatial context. Specifically, the representation coded in grid cells allows vector navigation to hidden goal locations. Recently, several studies have described non-spatial neuronal parallels of place and grid cells. This suggests that the coding principles giving rise to spatial place and grid cells may also support the general, arbitrary-topology cognitive maps Tolman originally envisaged. However, previously suggested grid cell-based mechanisms for vector navigation in 2D physical space do not generalise to arbitrary topologies.

In this chapter we contribute to a growing body of literature demonstrating the benefits of viewing physical space as a specific example of the more general concept of (RL) state-spaces. We replicate and add to work showing that grid-like firing patterns can be obtained as an efficient representation (principle component analysis) of the co-activity of place cells in 2D space, suggesting grid cells may represent the topology of the state space rather than 2D physical space per se. We then show how grid-like codes contain not only local information about co-active cells, but also readily available knowledge of approximate distances between every pair of locations (states) in the environment, allowing flexible and efficient navigation of environments with arbitrary topology. We refer to this navigation algorithm

as "intuitive planning". This chapter is the product of joint work carried out with Timothy H Muller.[1]

4.1 **Introduction**

It has been argued that in order to generate complex behaviour that is flexible to changes in circumstance and internal goals, animals maintain a model of the relationships between objects in the world, often termed a cognitive map (Tolman, 1948). In physical space the neural substrate for such a map is hypothesized to exist in the hippocampal formation, most notably in the form of place cells (O'Keefe and Nadel, 1978) and grid cells (Hafting et al., 2005).

Tolman's original concept of a cognitive map was not, however, restricted to spatial relationships, but encompassed relationships that subserved flexible behaviour across multiple domains of life (Tolman, 1948). Indeed, the hippocampal formation is critical for knowledge representation and generalisation in behavioural domains as distinct as social cognition (Sams-Dodd et al., 1997), reward learning (Gilboa et al., 2014), imagination (Hassabis et al., 2007) and certain types of memory (Nadel and Moscovitch, 1997). Furthermore, neurons coding for non-spatial information in a fashion analogous to place cells have been found in human (Quiroga et al., 2005) and rodent (Aronov et al., 2017b) hippocampus, and rodent entorhinal grid cells encode other continuous domains like time (Kraus et al., 2015) or auditory frequency space (Aronov et al., 2017b) with a grid-like code. Additionally, fMRI signatures of grid-like coding have been found in human entorhinal cortex (Bao et al., 2019; Constantinescu et al., 2016), as well as other regions.

Recently, we (Behrens et al., 2018b) and others (Dordek et al., 2016; Gustafson and Daw, 2011; Stachenfeld et al., 2017; Wilson et al., 2014) suggested a formal account of these representations, which views well-characterised spatial representations as a particular example of a more general coding framework. This framework, which relies on borrowing

[1] While most work was done collaboratively, this chapter includes a details explanation on the theoretical view of grid cells as a PCA of location information, while Tim's work demonstrated these principals in data from Caswell Barry's lab.

the formal tools of RL state-spaces, allows the abstract notion of a "cognitive map" to be formalized with the mathematics that describes general behavioural problems. Note that the key contribution of RL here is the reconceptualization of physical space as a state-space (formally, a Markov Decision Process or MDP, see section 1.3.2)), rather than optimising reward. As we will see below, this is achieved if place cells are formally thought of as encoding states in an RL world, i.e. if locations are considered as states (Gustafson and Daw, 2011; Stachenfeld et al., 2017).

By examining neural coding under this framework, it is possible to ask whether spatial representations observed in the brain may be explained as mathematically-defined representations (vectors/matrices) arising from the statistical/transition structure of 2D space. For example, place cell representations may be better explained as coding future or successive states (the "successor representation" (Dayan, 1993)) rather than simply the current state (Stachenfeld et al., 2017). Of particular relevance to us, which we explore further below, is the observation that grid cell firing may be thought of as the eigen-spectrum of local place cell transitions (Dordek et al., 2016; Stachenfeld et al., 2017) or other (non-linear) low-dimensional representations of 2D space (Banino et al., 2018). This suggests grid cell firing can be understood as encoding the relationships between locations/place cells and therefore summarising the statistics or structure of 2D space.

Storing an internal model of the structure of physical space has obvious advantages, such as facilitating flexible planning or inference of unobserved paths or links. However, it is not yet clear why it is advantageous to encode such a map both locally (using place cells that are active in specific locations) and globally (using grid cells that each fire in many different locations, but whose population can encode location precisely (Stemmler et al., 2015; Wei et al., 2015)). However, one attractive recent suggestion is that these grid codes can be used to efficiently solve navigation problems, as much of the necessary computation is already solved in the representation. Angles and distances between different points in the map can be extracted from grid codes using inexpensive computations that do not require searches through place cells (Banino et al., 2018; Bush et al., 2015; Stemmler et al., 2015), endowing the agent with a "sense of direction" for intuitively navigating the map. However, cognitive maps encoding non-spatial knowledge might have arbitrary topologies. Therefore, vector-based navigation that relies on the 2D-specific properties of grid cells will fail to navigate these maps. Hence it is important to establish how powerful grid-like representations may be used to flexibly and efficiently navigate arbitrary topology worlds as well as spatial worlds.

In this chapter we explore how casting spatial problems under a (RL) state-space framework may lead to insights about how representations emerge, as well as how these representations may be used to support flexible navigation in arbitrary topology worlds. First, we briefly explain what it means to cast spatial problems under an RL/state space framework. We then demonstrate how grid-like representations may be explained in this context as summarising the statistics of 2D space. To do this we reproduce the core result described in (Dordek et al., 2016) showing grid cells firing patterns emerge as a principle component analysis (PCA) of place cell activity, and explore this further in a couple of minor ways. Finally, unlike previous work focusing on the spatial domain (Bush et al., 2015; Stemmler et al., 2015), we show how once such powerful structural representations are stored, they can be used to perform flexible and efficient navigation in arbitrary topology as well as spatial environments (Baram et al., 2017).

4.2 The structure of 2D physical space

4.2.1 Casting space under a state space framework

We briefly describe how spatial coding can be formalised and unified with non-spatial coding under a state space framework by thinking of locations as states. Consider a simple linear track. By discretising the track, we can define the different locations along the track as different states, such that a state indexes a given location. If we assume a random walk policy, the relationships between states are defined by the $N_{states} \times N_{states}$ transition matrix T (or adjacency matrix A[1]). Element T_{ij} is the probability of transitioning from state i to j in the next time step: $T_{ij} = P(X^{t+1} = j \,|X^t = i)$. Hence, any non-zero elements must be neighbour states. This matrix therefore represents the topology of the environment.

[1] The adjacency and transition matrices are related, but their elements have different meanings. The adjacency matrix is a binary matrix (representing a graph), indicating which states (nodes) are connected (with and edge). Element A_{ij} of an adjacency matrix is either 1 or 0 depending on whether it is or isn't, respectively, *possible* to transition from state i state j to. The transition matrix elements are probabilities, and it is hence not necessarily a binary matrix. Element T_{ij} is the probability of transitioning from state i to j. The rows of T must therefore sum to 1.

P

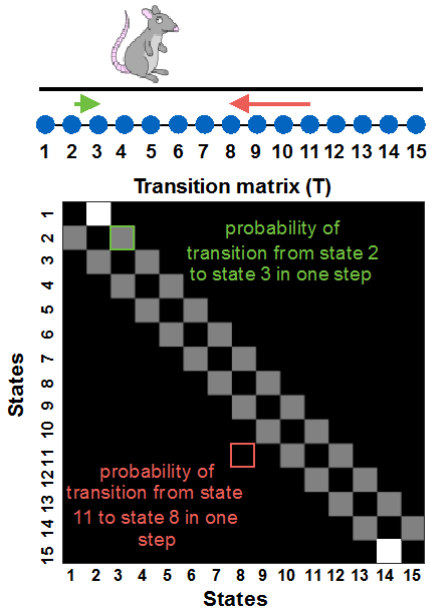


Figure 4.1 Casting physical space as a state space. Spatial cognition can be reconceptualised under a state space framework by thinking of (discretised) locations (or place cells) as states. For example, the locations on a linear track (top) can be defined as different states (blue dots). The relationships between these states can then be represented in a states-by-states transition matrix T such that T_{ij} is the probability of transitioning from state i to j. Here, transitions from a given state are only possible to the two adjacent states. Figure from (Behrens et al., 2018b).

4.2.2 The covariance structure of 2D physical space

For simplicity, we identify place cells with specific locations in space. While the true nature of place cells is much more complex (e.g. (Stachenfeld et al., 2017)), this assumption is useful as we can use the relationships between place cell activity as equivalent to the relationships between locations. In other words, the covariance (or co-activity) structure of our place cells is equivalent to the structure of continuous 2D space. This can be formalised as an $N_{cells} \times N_{cells}$ covariance matrix, where element (i, j) is defined as the covariance between place cell i and place cell j:

$$Cov_{i,j} = E[r_i - E[r_i]]E[r_j - E[r_j]]$$

Equation 4.1

Where E is the expectation across time, r_i is a vector (the length of the number of locations, $N_{loc} \geq N_{cells}$) of the firing rates of place cell i in each location. The covariance of two cells will be high if they two fire at similar times or equivalently when the agent/animal is at similar locations. This will occur when the cells have overlapping firing fields, i.e. they are indexing

physically close locations. Therefore, under a random walk policy, the covariance matrix is analogous to adjacency matrix, both summarising the structure of the environment.

Following (Dordek et al., 2016), we simulated the activity of place cells as an artificial agent performed a random walk in a square environment with periodic boundary conditions (i.e. a torus) to prevent boundary effects. We modelled place cells as simple Gaussians, randomly distributed in the simulated environment (**Figure 4.2a**). The firing rate r of place cell i at time t, when the agent is at location (x^t, y^t), is given by a 2D Gaussian distribution (Figure 4.2b):

$$r_i^t((x(t), y(t))) = exp\left(-\frac{(X(t) - C_i)^2}{2\sigma_i^2}\right)$$

Equation 4.2

where C_i is the receptive field centre of place cell i, and σ_i is its width (σ_i assumed to be the same in both dimensions). The agent's walk was random but with a bias such that it is not likely to turn with sharp angles and it is more likely to continue in roughly the current direction. As in (Dordek et al., 2016), the agent's movement is defined by:

$$D^{t+1} = modulo(D^t + \omega Z, 2\pi)$$

$$x^{t+1} = x^t + vcos(D^{t+1})$$

$$y^{t+1} = y^t + vsin(D^{t+1})$$

Equation 4.3

where D^t is the angle of the current direction, ω is the angular velocity, $Z \sim N(0,1)$ where N is the Gaussian distribution, i.e. Z induces randomness in the agent's direction but also a bias to continue travelling in the same direction (as the mean of N is 0), v is the linear velocity, and (x^t, y^t) is the location of the agent at timestep t. Boundary conditions for the agent's trajectory were also periodic (**Figure 4.2c**).

Through this simulation we constructed an $N_{cells} \times N_{time\ points}$ activity matrix of the firing rates (**Figure 4.2d**). Importantly, this matrix has been constructed only from local information: the (overlapping) activity of place cells in each time point. The agent was never shown the entire structure of the environment at once. Each pair of rows (cells) of this activity matrices was used as inputs to calculate the covariance between cells through

Equation 4.1, resulting in an $N_{cells} \times N_{cells}$ covariance matrix (**Figure 4.2e**), summarising the statistical structure of the toroidal 2D space.

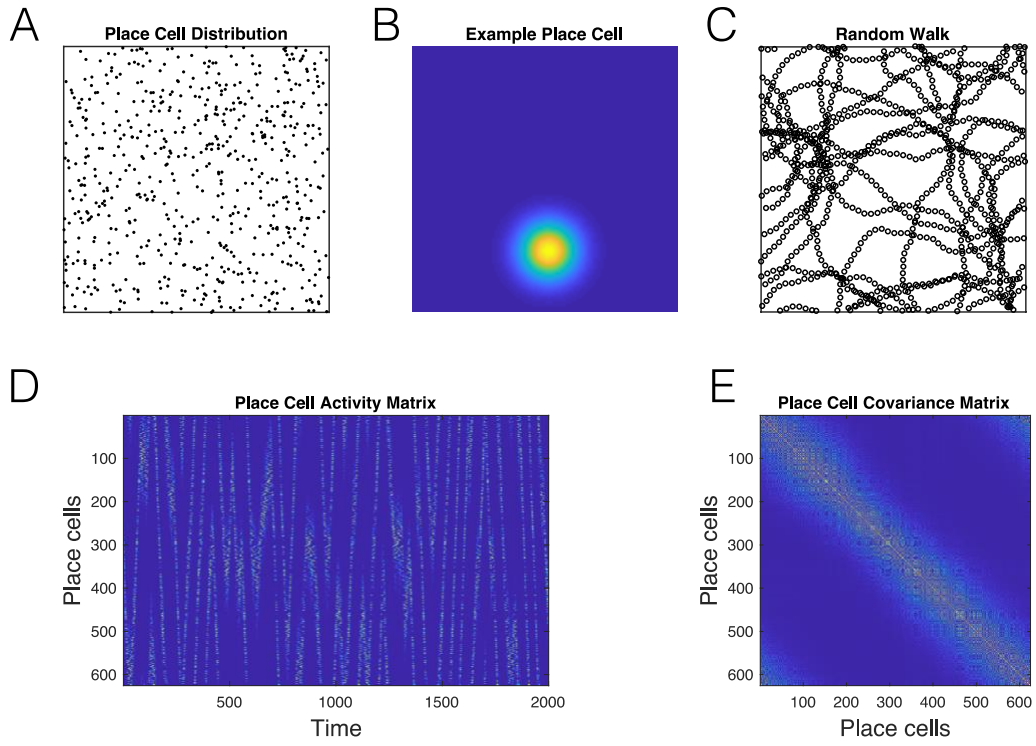


Figure 4.2 Constructing the place cell covariance from local observations. a. Uniform distribution of place cells (black dots denote place cell centres). **b.** An example place cell rate map. **c.** First 1000 timesteps of a random walk through the environment. Note the periodic boundary conditions. **d.** Activity of each place cell as a function of time in the first 2000 timesteps. **e.** Covariance matrix of the place cell activity matrix. This matrix summarises the structure of the square toroidal 2D space. Notably, we computed this matrix without ever having observed the structure, only local place cell activity over time.

4.2.3 Grid cells as an eigen-representation of the structure of 2D space

4.2.3.1 An intuitive explanation of Principal Component Analysis (PCA)

Another way of understanding why the covariance matrix is a good representation of the space spanned by the neurons (or more generally, any variable) is through Principal Component Analysis (PCA). In this data analysis method, a set of orthogonal components (vectors) is found, such that the first component accounts for as much of the variability in the data as possible. Equivalently, this component has the minimal square distance from all data points. The next component is the one that again accounts for the most variance, under the constraint that it has to be orthogonal to the first. In this way a set of N_{cells} orthogonal vectors is found. These components are the eigenvectors (EVs) of the covariance matrix of

the cells' activity, with the first component corresponding to the EV with the largest eigenvalue. This is not a coincidence: the covariance matrix, like any matrix, is a linear transformation, which transforms any N_{cells}-dimensional vector to another N_{cells}-dimensional vector. The EVs are vectors which maintain their direction under such transformation (they only change by a scalar – the eigenvalue). As the EVs form a basis for the N_{cells}-dimensional space, any N_{cells}-dimensional vector can be expressed as a linear combination of EVs. Hence, applying the transformation defined by the covariance matrix on any vector will result in a vector which is closer in its direction to the EV with the maximum eigenvector (how much closer it is to this vector than the other EVs depends on the ratio between eigenvalues). Doing this a large (infinite) number of times will result in convergence to the EV with the maximal eigenvector – i.e. to the principal component that has the minimal distance from the data points. So, in some sense, the covariance matrix transforms any N_{cells}-dimensional vector to be more similar to data actually observed from the original cells, i.e. to a vector that conforms to the structure of 2D space.

4.2.3.2 Principal components of place cell activity are grid-like

Following (Dordek et al., 2016), we show that the EVs of the covariance matrix of place cells activity, once projected back to 2D space, have grid-like properties. While we used MATLAB to calculate these EVs, (Dordek et al., 2016) go further and show this can also be calculated with a simple feed-forward neural network where place cells connect to output (grid) cells, using a learning rule that effectively performs PCA (Oja, 1982). The principal components are then represented in the synaptic weights between the place cells and the grid cell(s), such that a dot product between the place cells activity vector and the vector of their corresponding place-to-grid cell weights (the EV) results in a rate map of the grid cell activity in the 2D environment:

$$G(x,y)_k = \sum_{p=1}^{n} \varphi_p^k \, exp\left(-\frac{\left(x - c_{p,x}\right)^2}{2\sigma_x^2} - \frac{\left(y - c_{p,y}\right)^2}{2\sigma_y^2} \right)$$

Equation 4.4

where $G(x,y)_k$ is the firing rate of grid cell k at location (x,y), φ_p^k corresponds to the element of the kth eigenvector associated with place cell p, i.e. φ_p^k is the weight between place cell p and grid cell k (entry p in the kth eigenvector), and $c_{p,x}$ & $c_{p,y}$ are the (x,y) coordinates of the centre of place cell p. This view is consistent with the recently

acknowledged importance of the feedback connections from place cells to grid cells (Bonnevie et al., 2013; Langston et al., 2010; Wills et al., 2010). The resulting rate maps are grid-like, in that they are periodic over space (**Figure 4.3**). Note however, that the first few eigenvectors (those with the largest eigenvalues) have very low frequencies, hence are not obviously grid-like. For example, the first eigenvector just divides the environment to two halves.

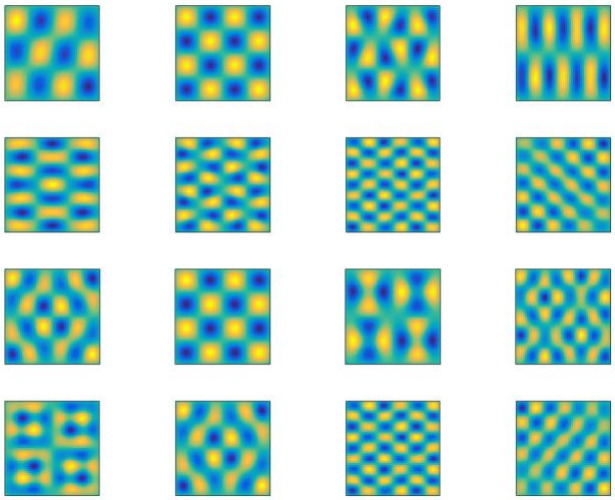


Figure 4.3 Grid firing fields forming a square lattice emerge from a square environment. Grid cell rate maps corresponding to the first 16 eigenvectors.

This periodic grid-like firing forms a square lattice over the environment, rather than a triangular/hexagonal lattice which grid cells are famous for (Hafting et al., 2005). (Dordek et al., 2016) found that a triangular lattice can be obtained by placing a non-negativity constraint on the PCA (or on the place to grid cell weights in the network implementation of the PCA), following the observation that place to grid cell weights are excitatory. We asked whether the non-negativity constraint is required for a triangular lattice, or whether other conditions may be sufficient. **Figure 4.4** shows that a triangular lattice is also obtained when the environment is defined by hexagonal rather than square boundary conditions, without a non-negativity constraint. This suggests that boundary conditions play an important part in defining grid fields from eigenvectors.

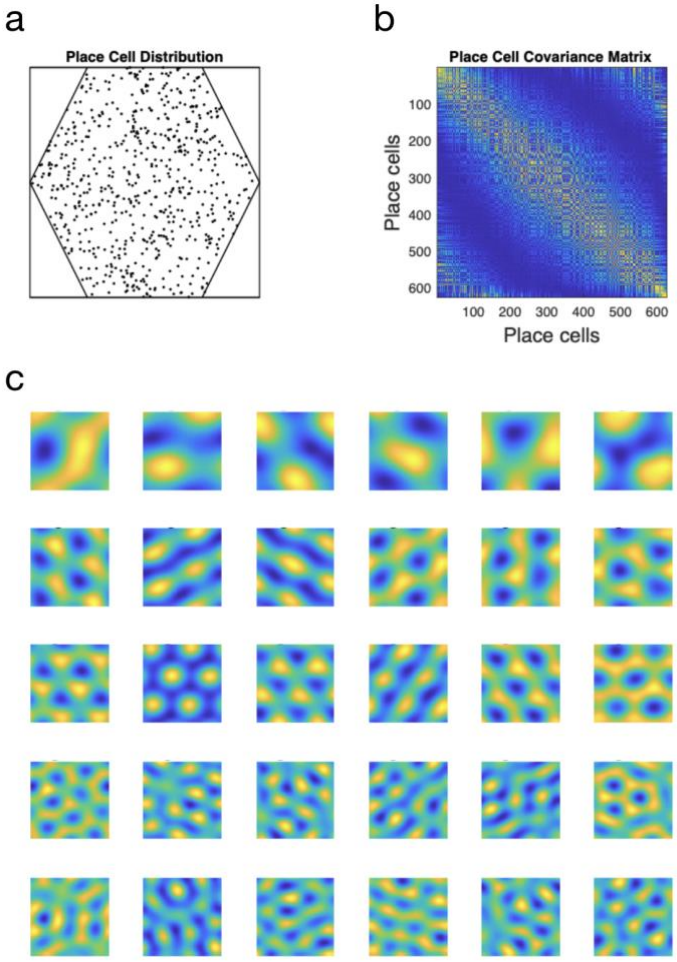


Figure 4.4 Grid firing fields forming a triangular lattice emerge from a hexagonal environment. a. Centres of place cells uniformly distributed in a world with hexagonal boundary conditions. **b.** The covariance matrix of place cell activity in the hexagonal world with periodic boundary conditions. **c.** First 30 eigenvectors.

4.3 **Intuitive planning: global navigation through cognitive maps using grid-like codes**

In the previous section we demonstrated how local observation, in the form of co-activity of neighbouring place cells, can result in global representations of the environment – cells with periodic lattice firing patterns, spanning the entire environment. In this section, we explore how such global representations can be useful. We take as a starting assumption that the brain holds a representation of the eigenvectors of the adjacency matrix defining the environment – in physical space this would be stored in the grid cells. We show these eigenvectors, once represented, can be used to easily calculate distributions over the future occupancy of all states, and provide an approximate distance between any pair of states in

the environment without any search. While the idea of constructing basis sets for solving MDPs through a spectral analysis of graph-defining matrices is not new (e.g. proto value functions (Mahadevan, 2005)), their possible neural implementation in grid cells is novel. These approximate distances provide a "sense of direction" – similar to the vectors to goals that can be readily calculated from grid cells in physical space (Banino et al., 2018; Bush et al., 2015; Stemmler et al., 2015). Importantly, these approximate distances can be calculated in non-spatial spaces, allowing for "intuitive" navigation (without any forward search or simulation) through maps with arbitrary topology. Hence, we show how through powerful global (grid) representations, learned only from local observations, an agent can perform "intuitive planning", regardless of the structure of the environment.

4.3.1 Global distances from local learning

Consider the adjacency matrix A, which defines the pairwise relationships between all states: element $A_{ij} = 1$ if it is possible to transition between state i and j, $A_{ij} = 0$ otherwise. We choose the adjacency matrix A over the transition matrix T since we borrow measures from spectral graph theory that are derived from A. Assuming a random walk policy, we can compute the distribution over future states n timesteps in the future by computing A^n (**Figure 4.5**). Element A^n_{ij} is the number of walks[1] of length n from states i to j. As A^n is simply a power of A, it shares the same eigenvectors of A:

$$A^n = \varphi \lambda^n \varphi^{-1}$$

Equation 4.5

where the columns of the matrix φ are the eigenvectors of A and λ is a diagonal matrix with the corresponding eigenvalues on the diagonal. λ^n is also a λ^n diagonal matrix. Furthermore, any summation of $a_i A^{n_i}$, (where n_i is an integer, and a_i is an arbitrary scalar) will share these eigenvectors, so if we choose an approximate distance metric that is a weighted sum of future states, it is easily computable from the eigenvectors of A.

[1] A walk between two nodes in a graph may revisit nodes and links along the way, whereas a path is a sequence of different nodes.

For example, consider the following two weighted sums:

$$\sum_{k=0}^{\infty} \frac{A^k}{k!} = e^A = \varphi e^{\lambda} \varphi^{-1}$$

Equation 4.6

and

$$\sum_{k=0}^{\infty} \gamma^k A^k = (I - \gamma A)^{-1} = \varphi (I - \gamma \lambda)^{-1} \varphi^{-1}$$

Equation 4.7

The dominant contribution to element (i, j) of these matrices is the shortest path between i and j (minimal k for which $A^k_{i,j}$ is not 0). This is because the contribution of longer paths (higher k) is dramatically downweighted due to the factorial or exponential decay. Hence element (i, j) contains information closely related to the shortest distance between the states. This is shown empirically below, including some exceptions.

In graph theory, these two summations (Equation 4.6 & Equation 4.7) are both used as the definition of the communicability between nodes of a graph (Estrada and Higham, 2010). The communicability between nodes i and j (C_{ij}, where C is the communicability matrix) will be higher if information flows (diffuses) more easily between nodes i and j. Communicability was proposed as a generalisation of the concept of the shortest path to include longer paths that may be relevant for information flow between pairs of nodes (Estrada et al., 2009). Notably, the resolvent summation is closely related to the successor representation in reinforcement learning (Dayan, 1993), which is computed from T rather than A.

Critically, all of these matrices can be easily computed from the eigenvectors and eigenvalues of A, without any simulation through all paths in the environment. In order to obtain these summations, one can simply recombine the eigenvectors of A with different weights (scalars in a diagonal matrix, Figure 4.5). Instead of the original eigenvalues λ, λ^n will give the distribution over states n steps in the future, e^{λ} will give the matrix exponent, and $(I - \gamma\lambda)^{-1}$ will give the resolvent. Furthermore, these reweightings are inexpensive to implement because the number of reweightings required is equal to the number of

eigenvectors (grid cells), rather than the number of relationships that can be computed. Hence, measures that approximate the shortest distance between each pair of states may be efficiently computed from grid-like cells.

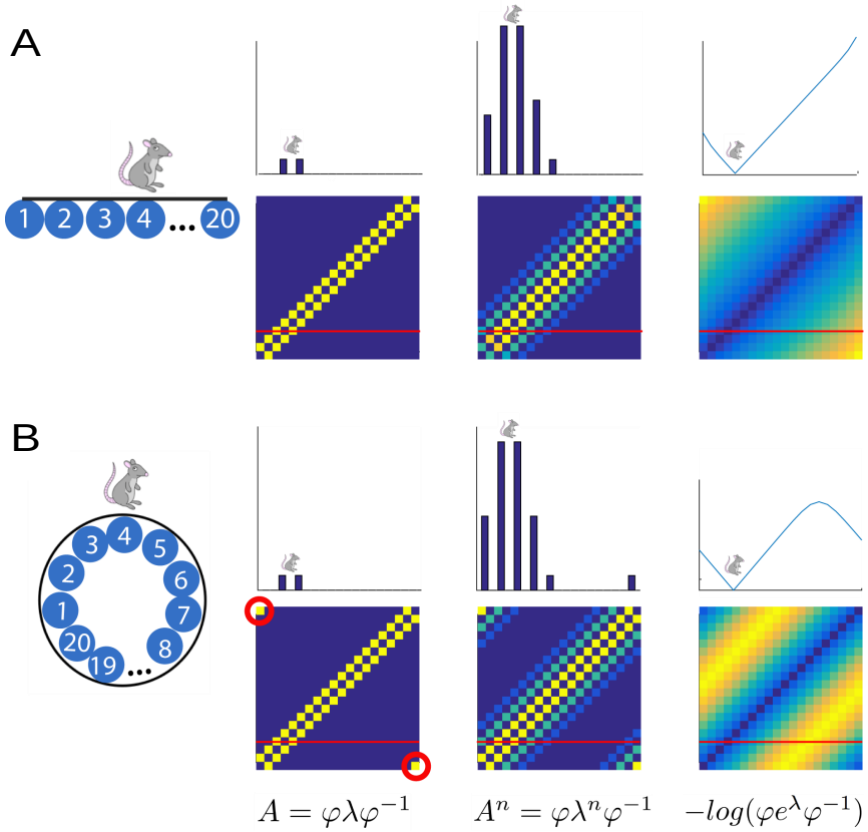


Figure 4.5 Construction of local and global knowledge from locally-learnt eigenvectors. a. A bidirectional linear track. For the three columns on the right, bottom is a matrix and top is a slice through that matrix (red line) at state four. From left to right: Schematic of the environment with a mouse occupying state four; the associated adjacency matrix A; A^n, where $n = 5$, and therefore A_{ij} is the number of possible walks of length n from state i to state j; $log(C)$, where C_{ij} is the communicability from state i to state j (communicability defined in the text). **b.** The same format as **a**. for a bidirectional circular track. Red circles in the adjacency matrix mark the new adjacency between the ends of the linear track in **a.**, turning it into a circular track. **Bottom**: Crucially, not only the matrix A containing local knowledge, but also A^n and $log(C)$ containing global knowledge can be constructed from the eigenvectors φ and eigenvalues λ of the matrix A. The local knowledge in state four is identical across the two environments (compare the slices through the adjacency matrices in **a.** and **b.**), but the global knowledge, obtained from the eigenvectors of A, is very different (compare slices in A^n and $log(C)$ across **a.** and **b.**).

We emphasise that these measures work for arbitrary topology environments. They accurately reflect communicability in environments that cannot be easily mapped on to 2D maps, such as those containing wormholes linking disparate parts of a map, as well as environments with one-way links between states, which result in complex eigenvectors and eigenvalues. Such arbitrary topologies are more likely those that are found in non-spatial cognitive maps.

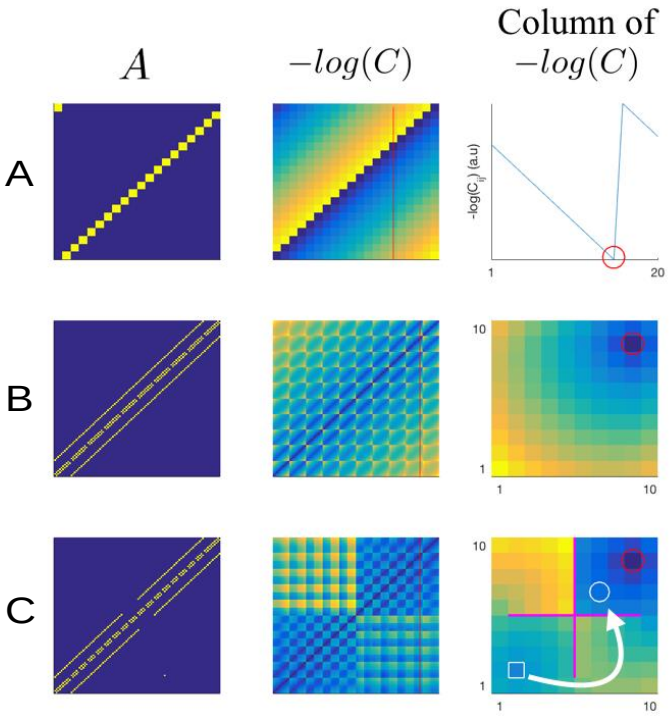


Figure 4.6 Communicability measures in arbitrary topologies. a. A one-way circular track. Left: The adjacency matrix A. Middle: the matrix $-log(C)$ where element C_{ij} is the communicability between states i and j. Red line marks the column of a target state (note this is different to **Figure 4.5** where it is the row of the current state). Right: This column represented on the space defined by A. Red circle indicates target state. **b.** A two dimensional 10x10 box. The same format as A except the target column of $-log(C)$ is represented as a heatmap on a two-dimensional environment with warmer colours corresponding to higher values. **c.** A box with barriers and a wormhole connecting states which are not adjacent in a two-dimensional space. Same format as b. States separated by a magenta line are not adjacent. White squares and circles mark entrance and exit gates of a wormhole, respectively: the states are adjacent yet distant in the regular two-dimensional topology. The conventions used in this figure will be used throughout the rest of the chapter.

Both the matrix exponential and resolvent measures can be approximately linearly related to the true distance, as defined by the length of the shortest path between states, by taking their $-log$ (**Figure 4.7**), a measure we refer to as *intuitive distance*. This relationship is dependent on the sets of walks between pairs of states with the same distance. In some environments, the set of walks between a pair of states can be obtained by a permutation from the set of walks between another pair of states with the same true distance. In such a case there is no variance in the relationship between intuitive and true distance (**Figure 4.7**a). However, in environments where there are many short, yet not necessarily shortest, paths between states, the communicability might differ between pairs of states with the same true distance, and hence so will the intuitive distance (**Figure 4.7**b). This predicts failures to navigate the shortest path between states (see below).

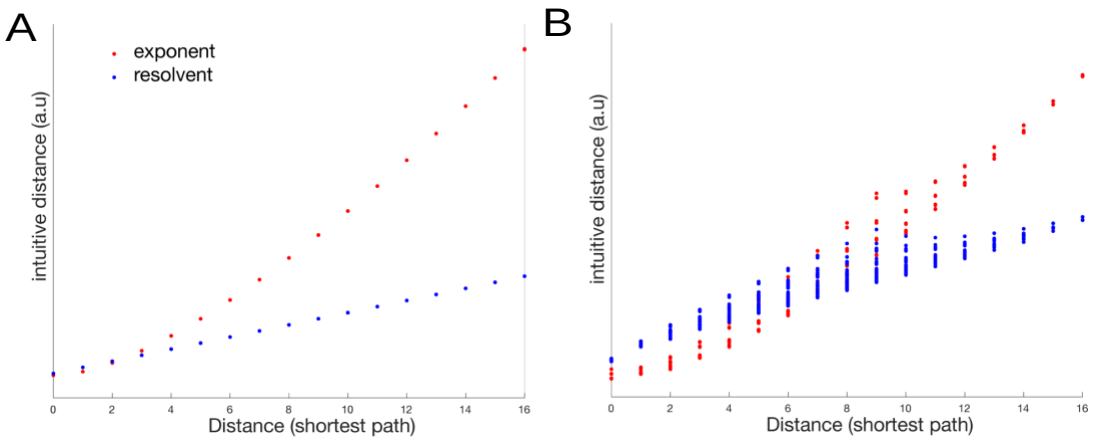


Figure 4.7 Approximately linear relationship between intuitive distance and true distance. Intuitive distances between pairs of states are approximately linearly related to the true distance, as defined by the length of the shortest path. **a.** Intuitive distance reliably increases with true distance on a bidirectional, one-dimensional track (the same as that in Figure 4.5**a**). **b.** Spread of intuitive distances for a given true distance in a 10x10 two-dimensional box (the same as that in **Figure 4.6b**), because communicability is influenced by the number of all possible walks between states.

4.3.2 Global navigation from local computation

Given current and target states, as well as the eigenvectors and eigenvalues, all calculations necessary for navigation are local: The agent needs to calculate which of the available next states has the shortest intuitive distance to the target. This can be done by calculating for target state T and each neighbouring state, N:

$$C_{NT} = \sum_{v=1}^{n} \varphi_v(N)\varphi_v(T)e^{\lambda_v} \text{ or } C_{NT} = \sum_{v=1}^{n} \varphi_v(N)\varphi_v(T)\,(1-\gamma\lambda_v)^{-1}$$

Equation 4.8

in the case of the matrix exponential or resolvent, respectively, where $\varphi_v(N)$ is the Nth element of the vth eigenvector of A (associated with eigenvalue λ_v), and n is the number of eigenvectors. The agent then simply proceeds to the neighbour state with the maximal C_{NT}, or equivalently minimal intuitive distance $-\log(C_{NT})$. This procedure finds trajectories that are approximately globally optimal without any global (tree) search. Because it stores the relational structure, the algorithm can flexibly reroute when the values of candidate target states change, for example due to changing internal goals.

Note that the "local" nature of the computation is only local insofar as the eigenvectors (grid cells) are already represented. Learning these eigenvectors is not a simple feat, and exposure to the entire transition matrix is needed (perhaps, in physical space through visual input). An intriguing possibility is that eigenvectors from well-known environments are "reused" in novel environments and thus bypass the need to explore the entire map. We explore ideas along these lines in the next two chapters.

4.3.3 Navigation using intuitive planning

We constructed several discrete-state mazes and tested whether intuitive planning could successfully navigate between start and target locations. We demonstrate that the algorithm successfully navigates the shortest route to the target in several two-dimensional, undirected mazes (**Figure 4.8**a-c). Cognitive maps encoding non-spatial relationships, however, will likely have arbitrary topologies. We show that the algorithm generalises to arbitrary topologies as its performance is not affected by the insertion of a wormhole into the maze, rendering it no longer topologically two-dimensional (**Figure 4.8**d-f). Successful navigation of the shortest route to the target, as defined by the minimum number of transitions to reach it, is achieved in all of the above mazes when using both the matrix exponential and resolvent measures. The mazes in **Figure 4.8**a-f can be mapped on to two-dimensional spaces for visualisation purposes. The algorithm also achieved successful navigation to the target in randomly generated, unweighted graphs, provided there was a route to the target (**Figure 4.8**g). Note that if, as in the case of a one-way worm-whole, the adjacency matrix is not diagonalisable, the eigen-decomposition is not straight forward, and will include imaginary eigenvectors.

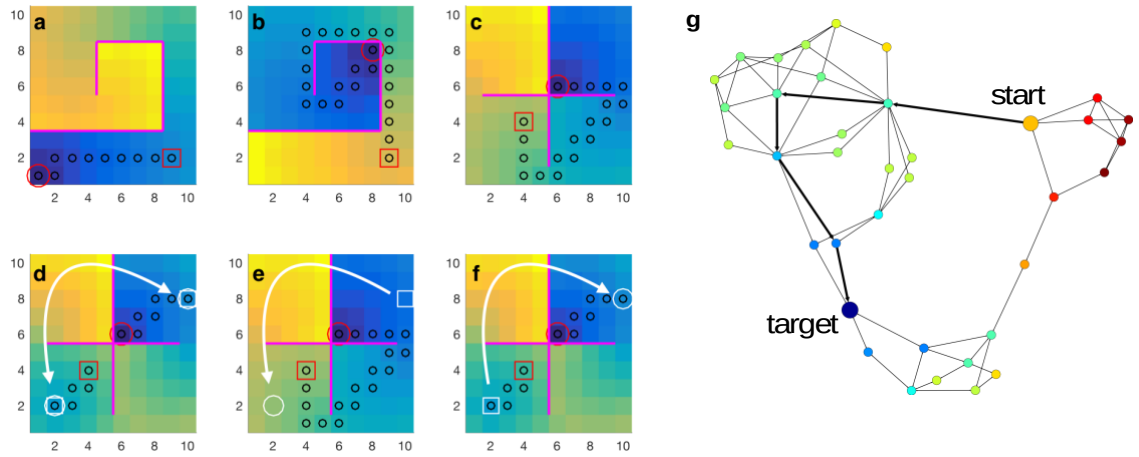


Figure 4.8 Navigation in arbitrary topology worlds. a-c. Successful navigation from start (red square) to target (red circle) states in two-dimensional worlds with barriers and differing start and target states. **d-f.** Introducing wormholes that connect disparate parts of the map and therefore warp it, rendering it no longer described by a simple map, does not affect the ability to successfully navigate to the target. **e-f.** Introducing one-way wormholes, resulting in complex eigenvalues and eigenvectors, does not affect the ability to successfully navigate to the target. In **a-f** navigation simulations, state transitions were permitted in the up/down and left/right directions (unless, for example, blocked by a barrier), and black circles denote the agent's path. **g.** Successful navigation in a graph of arbitrary topology. Colouring of nodes denotes intuitive distance (hotter is larger distance).

Since the relationship between intuitive distance and true distance is only approximately linear (**Figure 4.7**), it is possible to construct mazes where the algorithm successfully navigates to the target, but does not take the shortest route. This is achieved by associating states that are not part of the shortest route with multiple longer routes to the target, which contribute to the communicability measure. Such planning errors have interesting experimental predictions (**Figure 4.9**).

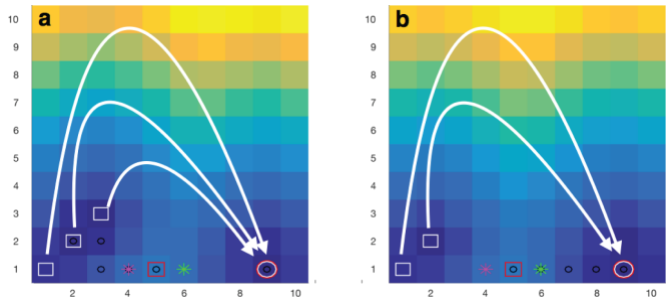


Figure 4.9 Failure to navigate the shortest path. Introducing multiple wormholes that cause multiple longer routes to the target to pass through a nearby state (magenta star) that is not in the direction of the shortest route (green star) causes the agent to take a longer route to the target (a). With fewer such wormholes, the shortest route is taken (b).

4.4 **Discussion**

In this chapter we have first introduced the notion of physical space as a specific relational structure, and its formalisation within the (RL) state-space framework. We then followed work by (Dordek et al., 2016) showing that the most efficient linear dimensionality reduction (PCA) of the relationships between locations (place cells) results in (eigen)vectors which have grid-like activity patterns, suggesting grid cells encode the relational structure dictated by the constraints of 2D space. Finally, we presented a novel navigation algorithm, which takes advantage of the power of the (grid) eigen-representations to endow agents with a "sense of direction" in environments with arbitrary topologies. This suggests a "sense of direction" navigational role of grid cells beyond 2D vector navigation .

It is important to note that the first two parts of the chapter are a replication of published work by (Dordek et al., 2016), except for a minor extension showing that conditions other than non-negativity (e.g. hexagonal boundary conditions) can result in hexagonal EVs (grid cells). I chose to include this replication in this thesis as the view of grid cells as an explicit representation of the structure of 2D space will be central to understanding the design and the interpretation of fMRI results in the next chapters.

Our emphasis here is not necessarily on the mechanistic interpretation of (Dordek et al., 2016) results as a place-to-grid cells neural circuit, but rather on the observation that grid cells encode the relational structure of 2D, in the sense that they are a low dimensional representation of the relations between locations. Such an embedding need not be linear as in the framework presented above: when a recurrent neural network is trained to perform path integration, grid-cells representations naturally emerge. Moreover, these representations are necessary and sufficient for the agent to generalise its learned navigation abilities to novel environments (Banino et al., 2018). Whether via a linear or a non-linear transformation, the view of grid cells as an efficient, low dimensional representation of 2D space is consistent with observations showing that the grid code requires the least number of neurons to encode an exact location (Wei et al., 2015) and that it is the most efficient way to tile 2D space (Mathis et al., 2014).

The place cells in (Dordek et al., 2016) and our simulations are purely allocentric and their activity does not reflect any effects of the animal's or agent's policy. As such, their covariance matrix, like an adjacency matrix, reflects the objective allocentric structure of the environment. However, several studies have shown that place cells seem to encode the environment in a policy-dependent manner, for example by over-representing rewarded locations (though see (Duvelle et al., 2019)) or skewing of place fields opposite to a preferred moving direction (Hollup et al., 2001). Beautiful theoretical work by (Stachenfeld et al., 2017) suggests that these observations can be explained by thinking of the place cells as encoding a prediction about the successor states/locations the animal will transition to – the successor representation (SR, Dayan, 1993). The SR-based agent can be thought of as halfway between a model-based and model-free agent. It has the flexibility of a model-based agent when encountered with a change of the reward function, but it does not necessitate computationally expensive forward simulations. The SR agent learns the transition structure of the environment on the fly through experience. However, the SR-agent is not flexible to changes in the transition structure and needs to relearn it through experience (akin to a

model-free agent in face of a change of the reward function). Interestingly, it has recently been demonstrated that human behaviour is consistent with these characteristics (Momennejad et al., 2016). As in the case of Gaussian place cells, an eigen-decomposition of the successor representation matrix results in grid-like eigenvectors (Stachenfeld et al., 2017). This strengthens the suggestion that grid cells indeed encode a low dimensional representation of the structure of 2D space – regardless of whether it is the "objective" 2D structure or a policy/experience-dependent subjective structure.

Our intuitive planning algorithm relies on a very similar successor-based representation – the communicability matrix. However, our work differs conceptually from the ideas of (Stachenfeld et al., 2017), as we suggest the communicability is stored in the grid cells (rather than the place cells), through the eigenvectors of the locally-learned adjacency/covariance matrix. This difference is useful as it enables place cells to retain a local, rather than global, scope. In another closely related work, Corneil & Gerstner (Corneil and Gerstner, 2015) have elegantly implemented a planning algorithm based on the eigen-decomposition of the SR of the environment in a recurrent attractor neural network. However, our observation that an SR-like representation can be represented directly in the grid cells with the eigenvectors is a novel one, though how exactly neurons can perform the suggested switching of weights between λ and e^{λ} is not clear. All successor-based navigational algorithms are as useful in arbitrary topologies as they are in 2D space.

The theoretical work presented in this chapter makes several experimental predictions. Intuitive planning predicts navigational failures when there are multiple paths to the goal associated with the states that are not part of the shortest path (**Figure 4.9**). In addition, interesting dissociations are predicted between scenarios where explicit planning (tree-search) are possible, and scenarios where such simulations will be too computationally expensive (where agents might use intuitive planning). For example, intuitive planning might not result in pre-play hippocampal activity (Johnson and Redish, 2007), while tree-search planning will. Behaviourally, when intuitive planning is used, the distance to the goal should not affect the difficulty of a task, while it dramatically increases the difficulty of a full tree-search. It is worth noting that there are indeed many cases when humans and animals plan without performing a full simulation of all possible future paths, for example through the use of hierarchies (Botvinick, 2012) or heuristics (Huys et al., 2015). An appropriate powerful representation can make expensive computation redundant (Akam et al., 2015). Such

powerful representations might be constructed offline, for example using replay (Liu et al., 2019; Wilson and McNaughton, 1994)

Finally, if grid cells are indeed the eigenvectors (or other low-dimensional projections) of the relationships between locations, it is possible that they will do the same for non-spatial environments, resulting in non-grid firing patterns that can be analytically predicted by eigenvectors. Such a view sees the griddiness of grid cells as a special case of a much wider phenomenon, resulting from the particular relational structure of 2D space. It is worth noting that we actually believe grid cells are not a perfect PCA of locations in space but rather a result of a more complex low-dimensional embedding. In such a case, even if the firing patterns of grid cells cannot easily be analytically predicted, we can still expect that grid cells (and perhaps other entorhinal cells taking part in the structural embedding) will have a more similar activity pattern in environments with the same relational structure than environments with a different structure. In the next chapters, we design a non-spatial RL task to test this prediction.

5 THE ENTORHINAL CORTEX ABSTRACTS AND GENERALISES THE STRUCTURE OF A REINFORCEMENT LEARNING TASK

Knowledge of the structure of a problem, such as relationships between stimuli, enables rapid learning and flexible inference. Humans and other animals can abstract this structural knowledge and generalise it to solve new problems. For example, in spatial reasoning, shortest-path inferences are immediate in new environments. Spatial structural transfer is thought to be mediated by grid cells in entorhinal cortex, which maintain their structure across different environments. In the previous chapter we demonstrated theoretical considerations supporting this claim. In this chapter, using fMRI, we show that entorhinal representations perform a much broader role in generalising the structure of problems. We introduce a task-remapping paradigm, where subjects solve multiple reinforcement learning (RL) problems differing in structural or sensory properties. We show that, as with physical space, entorhinal representations are preserved across different RL problems only if task structure is preserved.

5.1 Introduction

Reinforcement learning (RL) theory (Sutton and Barto, 1998) has given deep insights into the brain's algorithms for learning, but has remained relatively mute about the representations that are the foundation for this learning. For example, how might a task, or an element of a task, be represented in the brain? Some recent progress has been made through comparison with spatial navigation, where representations are better understood. It has been suggested that the same representations that map Euclidean space (such as hippocampal place cells (O'Keefe and Nadel, 1978) and entorhinal grid cells (Hafting et al., 2005)) may be extended to a broad range of non-spatial problems (Eichenbaum, 2017). In these cases, instead of representing physical location, they may represent location in an abstract space that captures the regularities of the task at hand (Behrens et al., 2018b; Garvert

et al., 2017; Gershman and Niv, 2010; Schuck et al., 2016; Stachenfeld et al., 2017; Wang et al., 2018; Whittington et al., 2019; Wilson et al., 2014).

One attractive corollary of these ideas is that non-spatial tasks might benefit from the profound representational efficiencies that are known in space. In space, as demonstrated in the last chapter, entorhinal cortex (EC) grid cells encode unique locations in the most efficient manner possible, given the correlation structure of observations on a 2D Euclidean plane (Banino et al., 2018; Dordek et al., 2016; Mathis et al., 2014; Stachenfeld et al., 2017). Furthermore, in remapping experiments, this representational structure is transferred across environments (Barry et al., 2012; Fyhn et al., 2007; Yoon et al., 2013). Hence, in new spatial environments, it is not necessary to re-learn the associations implied by 2D space. Instead it is sufficient to learn what sensory observation is where in the map (Whittington et al., 2019), and inferences (such as shortest paths (Baram et al., 2018; Bush et al., 2015; Tolman, 1948)) can be made immediately. The ability to transfer structural knowledge in non-spatial tasks would, similarly, bestow efficiencies. The structure of a problem, learnt in one situation, could be mapped onto a new situation with different sensory observations, and solutions could immediately be inferred. Behaviourally, it is clear that both humans and animals profit from such efficiencies. In psychology, this phenomenon is known as "learning-set" (Harlow, 1949) - subjects with prior exposure to the structure of a problem are routinely better at solving new examples.

Might similar mechanisms support both spatial and non-spatial generalisation? One prerequisite is that brain regions that contain grid representations in Euclidian spatial tasks should also represent the structure of a reinforcement learning task. A second is that these representations should (a) generalise across different sensory exemplars of the same structural problem and (b) differ between two problems of different structure.

We designed an RL task where we manipulated either the problem structure (by changing the correlation structure of serial bandits) or the sensory stimuli in a 2X2 factorial design. Using fMRI in humans, we found that the EC contained a representation that differed between different task structures, but generalised over different sensory examples of the same structure.

This design mirrors a spatial remapping experiment with two important differences: (1) It is non-spatial task elements that are being remapped, as opposed to locations in 2D space. (2) We include 2 separate structures as well as 2 separate environments (sensory stimuli) in

each structure. This design therefore enabled us to test for representations of the task structure, factorised from the representations of the stimuli they were tied to.

Using fMRI in humans, we found that the EC contained a representation that differed between different task structures, but generalised over different sensory examples of the same structure.

5.2 Methods

5.2.1 Subjects

We trained 49 volunteers over 4 days on an online version of the task. 17 subjects did not proceed to be scanned as they either did not comply with task demands (e.g. failed to complete training on time) or did not reach a behavioural criterion for knowledge of the outcome probabilities correlation structure (a difference of more than 0.3 between the fitted cross-term of the related stimuli and the mean of the fitted cross-terms of the unrelated stimuli, see below). 32 volunteers (aged 21-32 years, mean age 23.4, 18 females) with normal or corrected-to-normal vision and no history of neurological or psychiatric disorders participated in the fMRI experiment. 4 subjects were excluded from the analyses: 3 due to technical difficulties during the scanning, and one due to excessive motion. Hence, all analyses presented are based on data from 28 subjects. All subjects gave written informed consent and the study was approved by the University of Oxford ethics committee (reference: R51215/RE001).

5.2.2 Training and task

Subjects trained online for 4 days prior to the scan day, and were scanned on day 5. In each training session, subjects performed a task where three 1-armed bandits were interleaved pseudo-randomly. The bandits were cued by three different visual stimuli, randomly sampled without replacement for each session from a bank of 35 images. Two of the bandits (bandits A & B) had correlated outcome probabilities, while the third (C) was independent. There were two possible correlation structures for the outcome probabilities of bandits A & B: positive correlation (+Corr blocks) or negative correlation (-Corr blocks). Each training session comprised of 8 blocks with 60 trials each (3 stimuli X 20 trials per stimulus).

In each trial, subjects viewed one of the three stimuli and had to indicate their prediction for its associated binary outcome (a "good" or a "bad" outcome, demarked by a King of Hearts or Two of Spades card, respectively) by either accepting or rejecting the stimulus. Thus, there was always one correct answer in each trial: subjects should accept a stimulus if they predict the outcome to be the "good" outcome, and should reject if they predict the outcome to be the "bad" outcome (**Figure 5.1**a-b). Only in accepted trials, the subject would either win or lose a point depending on the outcome (i.e. accepting incorrectly resulted in loss of a point). Outcome identity was revealed in all trials, including rejection trials (except on training days 3 and 4, see below), even though the subject's score did not change in these trials (**Figure 5.1**b). Predictions of the outcomes could be formed based on the recent history of outcomes. The outcome probabilities switched pseudo-randomly between 0.9 and 0.1 with an average switch probability of 0.05 in each trial of the online training sessions. As the two correlated bandits switched together, subjects could use their knowledge of the correlation structure to learn from the outcome on stimulus A about stimulus B and vice versa (**Figure 5.1**c). Subjects were informed and reminded at the beginning of each training block that two of the bandits had correlated outcome probabilities, and that this correlation might be positive or negative. However, subjects had to infer which two bandits were correlated and what the sign of the correlation was.

The training schedule for an example subject is shown in Table 5.1. On day 1, subjects completed two sessions with a different triplet of stimuli used as the ABC cues in each session. In both sessions, two of the stimuli had a particular correlation structure, counterbalanced across subjects. That is, half of the subjects performed two sessions of 8 -Corr blocks each, while the other half performed two sessions of 8 +Corr blocks. Day 2 was identical to day 1, except that the two stimuli sets used were novel and the correlation structure was the one that the subject did not experience on day 1. On day 3, subjects again completed 2 sessions with a novel stimuli set per session, where the correlation structure between two of the stimuli alternated between blocks. The correlation structure was indexed by the background colour of stimuli (e.g**. Figure 5.1**c-d), and subjects were informed that the combination of stimuli set, background colour and correlation structure on day 3 would be the same in days 4 and 5, including in the scanning task. Thus, subjects could already learn the background colour-correlation structure mapping prior to entering the scanner. On day 4, subjects completed one session with all the 4 possible block types (2 stimuli sets x 2 correlation structures, e.g. **Figure 5.1**c-d). To reduce available information and facilitate

subjects' need to use the extra information afforded by the correlation structure, no counterfactual feedback was given on rejection trials in any of the last 15 trials of a block in training days 3 and 4.

Prior to scanning on day 5, subjects completed a pre-scanning reminder session with all 4 block types, again with the same stimuli set - background colour - correlation structure combinations as in days 3 and 4. In both the pre-scanning session and during scanning, full outcome feedback was given, including in all rejection trials.

During scanning, subjects completed 8 blocks of 30 trials each (10 trials per stimulus), with a break after 4 blocks for structural and field-maps scans. The two groups of 4 blocks included one of each of the 4 experimental block types in pseudo-random order, counterbalanced across subjects. Outcome probabilities switched faster in the scanner than during training due to the shorter blocks, with a switch probability of 0.15.

Before the first trial of each block, all three stimuli and the background colour of that block were presented. A trial was on average 11.5 seconds long, progressing in the following order (**Figure 5.1**a): 1) A stimulus would appear in the middle of the screen, together with the available choices, e.g. left for accept (corresponding to an index finger button press) and right to reject (middle finger button press). The left/right mapping to choices was counterbalanced across subjects but stable within-subject. 2) After 1.5 seconds, the frame of the stimulus and the accept/reject text turned white, indicating that choice can now be made. Subjects who indicated their choice prior to the appearance of the white frame lost half a point. 3) Subjects indicated their choice by pressing either the index or middle finger buttons. Next, a red rectangle appeared around the chosen option for 0.5 seconds. 4) A white fixation cross appeared for a variable period, drawn from an exponential distribution with a mean of 4.5 seconds and truncated between 3.5-5.5 seconds. The purpose of this long delay between choice and outcome was to enable the independent analysis of both periods, due to the sluggish nature of the hemodynamic response function. 5) The outcome of the trial appeared for 1 second in the middle of the screen - either the "good" outcome (King of Hearts card) or the "bad" outcome (Two of Spades card). If the subject has accepted the trial earlier, they will either win or lose a point: if they accepted correctly (outcome was "good"), a "sack of gold" image would appear on the left side of the screen to indicate that a point was gained; if they accepted incorrectly (outcome was "bad"), a "no sack of gold" image would appear to indicate that a point was lost. If the subject rejected the trial, no points would be won or lost (and hence no "sack of gold" or "no sack of gold" image would appear), but the outcome

card image would still appear (**Figure 5.1**b). Hence subjects received full counterfactual feedback. Note that subjects should reject trials if they predict the outcome to be the "bad" outcome, as they will lose a point for incorrectly accepting a trial. 6) A white fixation cross appeared for a variable inter-trial interval, drawn from an exponential distribution with a mean of 3 seconds and truncated between 2.5-4 seconds.

5.2.3 Behaviour modelling

We modelled the behaviour of the subjects using an adapted delta-rule model (Rescorla and Wagner, 1972). The original delta-rule model estimates the outcome probabilities for a given stimulus using the following equation:

$$\hat{g}^{S}{}_{t+1} = \hat{g}^{S}{}_{t} + \alpha\epsilon^{S}{}_{t}\,;\ \epsilon^{S}{}_{t} = y_t - \hat{g}^{S}{}_{t}$$

Equation 5.1

where $S \in \{A, B, C\}$; $\{A, B, C\}$ are the three stimuli presented in a block, $\hat{g}^{S}{}_{t}$ is the value estimation of stimulus S before trial t (which can be thought of as the "good" outcome probability estimation for stimulus S before trial t, only on the [-1,1] interval instead of [0,1]), α is the learning rate, ϵ is the outcome prediction error, and y_t is the outcome at trial t: $y_t = 1$ for the "good" outcome and $y_t = -1$ for the "bad" outcome. Note that the model estimates the stimulus value, based on the outcome identity (which card was obtained - not the reward actually obtained by the subject), and is agnostic to the choice the subject made. The stimulus value should be distinguished from the "chosen action value" used in GLM2 (see section 6.2), which has the same magnitude as the stimulus value, but has an opposite sign on rejection trials: in a trial where the subject's hypothetical estimate of the stimulus value is very low (close to -1), they will be confident in rejecting the trial, making the value of the chosen "reject" action high (close to $+1$). Similarly, ϵ should not be confused with the "correctness" prediction error used in GLM3 (see section 6.2) or GLM5 (section 6.3.2 and **Figure 6.4**), which has the same magnitude as ϵ but with a sign determined by the congruence between the subject's choice and the outcome. The stimulus value estimation can then be used by a "selector model" to make a choice in the next trial, by using a simple sigmoidal function:

$$P(choice\ on\ stimulus\ S\ on\ trial\ t+1 = accept) = (1 + e^{-\beta(\hat{g}^{S}{}_{t})})^{-1},$$

Equation 5.2

where β is the inverse temperature, controlling the randomness of the choice. In this model, α and β are free parameters that we fit to subjects' behaviour.

However, this model does not use any knowledge about the correlations between the outcome probabilities of different stimuli. To allow for this, we added three free parameters to the model, which we refer to as cross-terms. These parameters determine how information on one stimulus affects the outcome estimate of another stimulus.

Following an outcome on stimulus A, we update the outcome estimates of all three stimuli:

$$\hat{g}^{A}{}_{t+1} = (1 - \alpha)\hat{g}^{A}{}_{t} + \alpha y_t$$

$$\hat{g}^{B}{}_{t+1} = (1 - \alpha|H_{AB}|)\hat{g}^{B}{}_{t} + \alpha H_{AB} y_t$$

$$\hat{g}^{C}{}_{t+1} = (1 - \alpha|H_{AC}|)\hat{g}^{C}{}_{t} + \alpha H_{AC} y_t$$

Equation 5.3

where $-1 \leq H_{XY} \leq 1$ is the cross-term for stimuli X and Y. Note that the first equation (update of the estimate for stimulus A following an outcome on stimulus A) is identical to the update in the original delta-rule model. $H_{XY} = 1$ means stimuli X and Y are treated as the same: the outcome estimates for both stimuli will be updated in exactly the same way following feedback on one of them. $H_{XY} = -1$ means stimuli X and Y are treated as having opposite (anti-correlated) outcome probabilities. $H_{XY} = 0$ means the two outcome probabilities are treated as uncorrelated: the outcome estimates for stimulus X will not change following feedback on stimulus Y, and vice versa. Analogous updates occur when feedback is given on stimuli B or C.

The 5 free parameters of the model were fitted to the behaviour of each subject, separately for each of the 4 block types. Due to the small number of trials of a particular stimulus in a particular block type, we pooled each subject's data from the scanning session, the pre-scanning session and the last online session (training day 4, after subjects had already learned the mapping between stimuli set, background colour and correlation structure), and fitted the parameters separately for +Corr and -Corr blocks. This resulted in a total of 150 trials for each of the 4 block types. The outcome estimates for all stimuli were reset to 0 at the beginning of each block. We fitted the parameters by maximising the negative log likelihood of the data with respect to the parameters using the Matlab function *fmincon*. The learning

rate was constrained to be between 0 and 1, the cross-terms between -1 and 1, and the inverse temperature between 0 and 8. Finally, we tested whether subjects indeed used the relational structure correctly by averaging the cross-terms from the same relational structure within-subject (+Corr or -Corr, **Figure 5.2**) and performing a paired t-test between the AB cross-term and the mean of AC and BC cross-terms.

5.2.4 fMRI data acquisition and preprocessing

Data were acquired on a 3T Siemens Prisma scanner, using a 32-channel head coil. Functional scans were collected using a T2*-weighted echo-planar imaging (EPI) sequence with a multi-band acceleration factor of 3, within-plane acceleration factor (iPAT) of 2, TR = 1.235 s, TE = 20 ms, flip angle = 65 degrees, voxel resolution of 2x2x2 mm and a tilt of 30 degrees relative to axial axis. A field map with dual echo-time images (TE1 = 4.92ms, TE2 = 7.38ms, whole-brain coverage, voxel size 3x3x3 mm) was acquired to correct for geometric distortions. Structural scans were acquired using a T1-weighted MP-RAGE sequence with 1x1x1 mm voxels. Preprocessing was performed using tools from the fMRI Expert Analysis Tool (FEAT), part of FMRIB's Software Library (FSL (Jenkinson et al., 2012)). Data for each of the 8 blocks were preprocessed separately. Each block was aligned to the first, presaturated image using the motion-correction tool MCFLIRT (Jenkinson et al., 2002). Brain extraction was performed using automated brain extraction tool BET (Smith, 2002). All data were high-pass temporally filtered with a cut-off of 100 seconds. Registration of EPI images to high-resolution structural images and to standard (MNI) space was performed using FMRIB's Linear and Non-Linear Registration Tool (FLIRT and FNIRT (Jenkinson et al., 2002; Jenkinson and Smith, 2001)), respectively. The registration transformations were then used to move each block's EPI data to the native structural space, downsampled to 2x2x2 resolution. No spatial smoothing was performed during preprocessing (see below for different smoothing protocols for univariate and multivariate analyses).

5.2.5 RSA analyses

5.2.5.1 General description for all RSA analyses in this thesis

In this section, we first outline the steps common to all RSA analyses (Kriegeskorte et al., 2008), including those presented in the next chapter, and then describe details which were specific to the analysis presented in this chapter. All RSA analyses were conducted as follows:

1) A searchlight (Kriegeskorte et al., 2006) was constructed around each cortical voxel, including the 100 cortical voxels with the smallest surface-wise geodesic distance from the central voxel. This was performed using adapted scripts from the RSA toolbox (Nili et al., 2014 original surface-based searchlight scripts written by Joern Diedrichsen & Naveed Ejaz, code available at https://github.com/rsagroup/rsatoolbox) . The searchlight definition depends on cortical reconstruction and alignment performed via Freesurfer's *recon-all* command (Dale et al., 1999; Desikan et al., 2006; Fischl et al., 2002, 2001, 1999; Reuter et al., 2010; Ségonne et al., 2004). 2) For each searchlight, first-level univariate GLMs on unsmoothed data were conducted using the RSA toolbox, based on SPM12. Regression coefficients were then spatially pre-whitened within the searchlight using the RSA toolbox. 3) A distance metric was defined to summarise the representational geometry between conditions. The metrics used were either the cross-run correlation distance or a within-block correlation distance (See below). This resulted in data RDMs of size [number of conditions] x [number of conditions]. 4) A hypothesis about the representational geometry was formalised as a contrast between the mean of RDM elements that should "dissimilar" to each other and the mean of RDM elements that should be "similar" to each other. This contrast was calculated in each searchlight, resulting in a single contrast map per subject. 5) The contrast maps of the different subjects were aligned on the common cortical surface using Freesurfer-based scripts adapted from the RSA toolbox. 6) Group-level statistical significance of the contrast (equivalent to a one-sided t-test) and family-wise error (FWE) correction were performed using permutation tests (Nichols and Holmes, 2002) using PALM (Winkler et al., 2014). In this procedure, the contrast values from each subject were randomly multiplied by either 1 or -1, following the null hypothesis that the contrasts are symmetric around 0. The test's statistic was then defined as the cross-subject average of contrast values. This was repeated 10000 times, creating a null distribution of the means. The true value of this mean was then compared to this null distribution. The resulting (uncorrected) p-value map is displayed in **Figure 5.3**d as a heat map, at a threshold of $P<0.01$. To correct for multiple comparisons, we thresholded the (uncorrected) P-map at $P<0.001$, and measured the mass of all surviving clusters. We then repeated this procedure for each of the 10000 random sign-flip iterations described above, and created a null distribution of cluster masses by saving only the mass of the largest cluster in each iteration. Comparing the true cluster masses to the resulting null distributions results in FWE-corrected P-values.

5.2.5.2 Plotting

The paired mean difference across subjects between the two groups of RDM elements at particular vertices of interest is visualised in the Gardner-Altman estimation plots in figure **Figure 5.3**c. The figures were generated using an adaptation of the openly available Matlab package DABEST (Ho et al., 2019). Note that these plots containing data from peaks of FWE-surviving clusters, which are subject to selection bias (and are shown for visualisation purposes only; **Figure 5.3**c).

5.2.5.3 First-level GLM details

In the following descriptions, [A, B, C] refer to the three stimuli presented in a particular block, where the outcome probabilities associated with A & B were correlated (either positively or negatively, depending on the block).

To search for a representation of the relational structure between stimuli in the task (**Figure 5.3 & Figure 5.4**), we conducted a GLM (GLM1) which included the following regressors per each block: 3 main effect regressors ([A],[B],[C]) modelling the times of stimulus presentation; 3 main effect regressors ([A], [B], [C]) modelling the times of outcome presentation; 2 regressors modelling button presses across all stimuli: one modelling all "accept" trials and one modelling all "reject" trials; 6 motion parameters as nuisance regressors; bias term modelling the mean activity in each block. Only the 2 regressors modelling the presentation of the two related stimuli (A&B) in each block were used in the "relational structure representation" analysis (as stimulus C was not part of a relational structure, **Figure 5.3** top and **Figure 5.4**), while all stimulus presentation regressors (A, B & C) were used in the "visual stimulus identity representation" analysis (**Figure 5.3**, bottom).

5.2.5.4 Cross-run correlation distance

Each condition (a particular stimulus under a particular structure) had two patterns (100-long vectors of spatially pre-whitened regression coefficients) – one from each independent run. We defined the cross-run correlation distance between each pair of conditions by averaging the following quantities: [correlation distance between condition i pattern from run 1 and condition j pattern from run 2] and [correlation distance between condition i pattern from run 2 and condition j pattern from run 1]. This resulted in [number of conditions] x [number of conditions] data RDM for each searchlight (8x8 for the "relational structure representation" analysis; 12x12 for the "visual stimulus identity representation" analysis).

5.2.5.5 Hypothesis testing and group analysis

We then defined a hypothesis-driven contrast between RDM elements: In the main "relational structure representation" analysis, "different structure" elements should be more dissimilar to each than "same structure" elements. For the control analyses in **Figure 5.4**, we ignored elements of the same visual stimulus (**Figure 5.4**a) or the same stimuli set (**Figure 5.4**b). In the "visual stimulus identity representation" analysis, "different visual stimulus" elements should be more dissimilar than "same visual stimulus" elements. Using the maximum cluster mass statistic (Nichols and Holmes, 2002) for multiple comparisons correction, we report clusters that survived FWE correction at a cluster-forming threshold of $P<0.001$ within a cortical hemisphere (green clusters in **Figure 5.3** and **Figure 5.4**).

5.3 **Results**

Full description of the task and training protocols can be found in section 5.2.2. Briefly, subjects performed a task where three 1-armed bandits were interleaved pseudo-randomly. Two of the bandits (bandits A & B) had correlated outcome probabilities, while the third (C) was independent. Crucially, we manipulated two features of the task across the different blocks: 1) the sign of the correlation between the A & B bandit probabilities (+Corr or -Corr). We refer to this correlation as the relational structure of the stimuli. 2) The stimuli set, with two possible triplets of images. Thus, in the scanning session there were 4 block types, each with a specific combination of a relational structure and a stimuli set, arranged in a 2x2 factorial design (**Figure 5.1**d). The fMRI experiment comprised of 8 blocks, divided into two independent runs of each of the four block-types, with a pseudo-random block order counterbalanced across subjects. Hence, in total, subjects completed 8 scanning blocks. As the two correlated bandits switched together (**Figure 5.1**c), subjects could use their knowledge of the correlation structure to learn from the outcome on one related stimulus about the other.

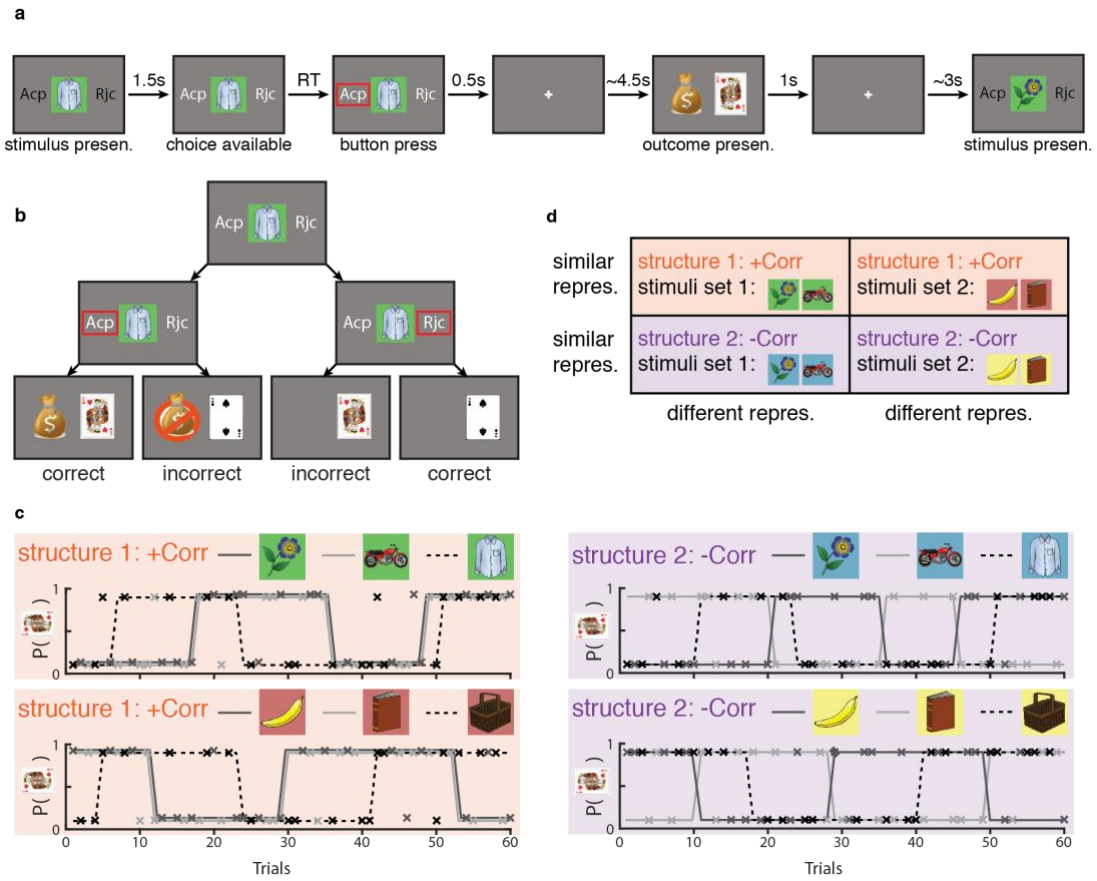


Figure 5.1 Task description. a. Timeline of an example scanning trial. **b.** Possible progressions of a single scanning trial. **c.** Example of the reward schedule for one subject. Solid grey lines and dashed black line are the probabilities of a good outcome for the related stimuli and the control stimulus, respectively. Xs mark the stimulus (colour) and actual binary outcome (y-axis: 0.1 and 0.9 are bad and good outcomes, respectively) in each trial. For visualisation purposes, the two 30 trials long scanning blocks of each of the four conditions were concatenated. While related stimuli in +Corr blocks are associated with exactly the same probability, their corresponding light and dark grey lines are slightly offset for visualisation purposes. **d.** Experimental design and neural predictions: 2x2 factorial design of stimuli set x relational structure.

5.3.1 Behavioural modelling and subject behaviour

Full description of the behavioural modelling is in section 5.2.3. We modelled the subjects' behaviour using an adapted delta-rule, with a crucial addition of "cross-terms" that enable learning from one stimulus to another. Note this is not intended to be a process model of how the brain solves the task. It is intended as a descriptive model that allows us to test whether, and to what extent, subjects are guided by the relational structure of the task. The model separately tracks the probabilities of a "good" outcome associated with the three stimuli, all of which are updated every trial. Following an outcome at trial t where stimulus X was presented, the estimate of the outcome probability associated with X is updated according to the classic delta-rule

T

$$\hat{g}^{X}{}_{t+1} = (1 - \alpha)\hat{g}^{X}{}_{t} + \alpha y_t$$

where $\hat{g}^{X}{}_{t}$ is the outcome probability estimation for stimulus X before trial t, α is the learning rate, and $y_t \in \{-1,1\}$ is the binary outcome at trial t.

Crucially, the estimates of the probabilities associated with the two other stimuli are also updated:

$$\hat{g}^{Y}{}_{t+1} = (1 - \alpha|H_{XY}|)\hat{g}^{Y}{}_{t} + \alpha H_{XY} y_t$$

Where H_{XY} is the cross-term between stimuli X and Y, fitted to the subject's behaviour in each block. $H_{XY} = 1$ indicates subjects treated X and Y as the same stimulus for learning purposes – for +Corr stimuli pairs this is the correct correlation structure. Similarly, $H_{XY} = -1$ and $H_{XY} = 0$ indicates correct correlation knowledge for -Corr and 0Corr pairs, respectively. y_t is still the outcome in trial t (a trial where stimulus X was presented).

We henceforth refer to the full model, including the 3 cross-terms, as STRUCT. We also used a structure-naïve Rescorla-Wagner model (equivalent to setting the cross-terms in the STRUCT model to 0), which we refer to as NAÏVE.

To determine whether subjects used the relational structure to inform their decisions we employed two approaches: a cross-validation approach and a cross-terms analysis approach. In the cross-validation analysis, for each subject we separately fitted the parameters of the STRUCT and NAÏVE models to data from the 4 block-types, treating trials from each block-type as a separate training set. This resulted in 8 fitted models: 4 STRUCT models and 4 NAÏVE models. We then performed cross-validation by testing how well each of the 8 models predicts subjects' choices in all 4 datasets (Figure 5.2a). As expected, STRUCT models generalised much better when trained and tested on different datasets of the same relational structure (despite different stimuli, squares highlighted in pink), than when trained on one relational structure and tested on the other (squares highlighted in green; comparison of sum(-log(likelihood)) of pink vs green elements; two-tailed paired t-test, t(27) =9.54, P<10^-9). Notably, the within-structure cross-validated STRUCT models (squares highlighted in pink) performed better than the NAÏVE models, even when the latter were trained and tested on the same data (squares highlighted in grey; compare pink vs grey elements in Figure 2a: two-tailed paired t-test, t(27) =4.29, P=0.0002). To further demonstrate this effect, we plotted the histograms of outcome probability estimations ($\hat{g}$)

from the relevant within-structure cross-validated STRUCT model (pink elements in Figure 5.2a; Figure 5.2b, left) and the NAÏVE model trained and tested on the same data (grey elements in Figure 5.2a; Figure 2b, right), split by subjects' choices. When the models made different predictions, subjects tended to choose in accordance with the within-structure cross-validated STRUCT models (Figure 5.2b). These results suggest that sensitivity to the correlation structure, here afforded by the cross-terms of STRUCT models, is needed to capture subjects' behaviour.

Due to the small number of trials in each block-type, for the analyses henceforth we trained STRUCT and NAÏVE on each subject's concatenated data from all blocks of the same structure. This resulted in 10 (5 parameters x 2 structures) and 4 (2 parameters x 2 structures) fitted parameters per subject for the STRUCT and NAÏVE models, respectively. The fitted cross-terms of the STRUCT model indicated that subjects indeed used the correlation structure correctly (Figure 5.2c; Corr vs 0Corr (mean across all blocks of $\left|H_{AB}{}^{-,+}\right|$ vs $(|H_{AC}| + |H_{BC}|)/2$), one-tailed paired t-test, t(27) = 13.06, P<10^-12). Subjects benefited from the extra information afforded by the correlation structure, and performed better in trials of one of the two related stimuli than in the control stimulus trials (Figure 5.2d).

Note that as we have excluded subjects who did not pass a behavioural threshold from the experiment, it is not possible to draw conclusions from these results to the general population. Our main purpose in presenting the behavioural results is for the interpretation of the neural signals: subjects who underwent scanning indeed knew and used the relational structure.

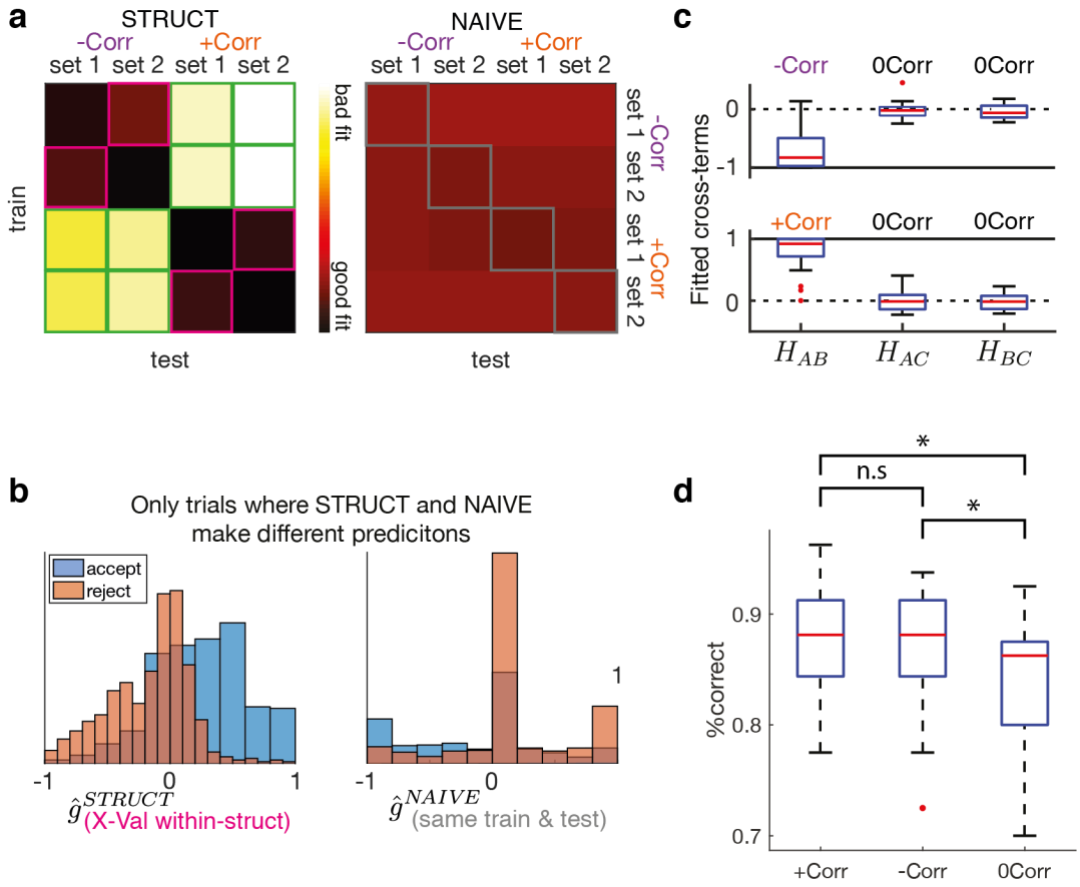


Figure 5.2 Subjects behaviour. a. Negative log likelihoods for STRUCT (left) and NAÏVE (right) models (same scale for both matrices). Pink elements: STRUCT models, cross-validated within-structure. Green elements: STRUCT models, cross-validated across structures. Grey elements: NAÏVE models, trained and tested on the same data. **b.** Histograms of the estimated outcome probabilities for trials where subjects accepted (blue) or rejected (orange). Left: STRUCT models trained on data with the same structure but different stimuli set (pink elements in **a.**). Right: NAÏVE models, trained and tested on the same data (grey elements in **a.**). Histograms only include trials where the models make different predictions. **c.** Fitted cross-terms for pairs of stimuli in all -Corr (top) and +Corr (bottom) blocks. Red central line is the median, the box edges are the 25th and 75th percentiles, the whiskers extend to the most extreme datapoints that are not considered outliers, and the outliers are plotted as red circles. **d.** The STRUCT model is more accurate in capturing subjects' behaviour in trials of related stimuli than trials of the control stimulus (two-tailed paired t-tests: -Corr vs 0Corr $t(27)=2.66$ $P=0.013$; +Corr vs 0Corr paired $t(27)=3.02$ $P=0.005$, +Corr vs -Corr paired $t(27)=8.3$ $P>0.4$).

5.3.2 Entorhinal representations generalise across tasks with the same structure but not those with different structures.

A representation of the relational structure of the task should be similar (generalise) for stimuli which are part of the same relational structure, but dissimilar for stimuli under a different relational structure. We asked whether any region on the cortical surface displayed these properties at the times of stimulus presentation, using Representational Similarity Analysis (RSA (Kriegeskorte et al., 2008)) with a searchlight approach.

A detailed description of the RSA analyses can be found in section 5.2.5. Briefly, a searchlight centred on a cortical voxel consisted of the 100 surrounding voxels with the smallest surface-wise geodesic distance from the central voxel. For each searchlight, we obtained 16 patterns of (whitened within-searchlight) regression coefficients of the responses to presentations of each of the two related stimuli (A and B) in each of the 8 blocks (GLM1). In other words, we obtained two patterns, one from each of the repeated runs, for each of our 8 experimental conditions (a particular stimulus under a particular correlation structure). To control for effects of time, we used a "cross-run correlation distance" where only patterns from different runs (i.e. more than 30 minutes apart) were correlated with each other. That is, to define the distance $d_{i,j}$ between conditions i and j we first calculated the correlation distance $(1 - r)$ between the condition i pattern from run 1 and condition j pattern from run 2, and then calculated the correlation distance between the condition j pattern from run 1 and condition i pattern from run 2. $d_{i,j}$ was defined as the mean of these two distances. Notably, this means that the diagonal in the symmetric Representational Dissimilarity Matrix (RDM) is meaningful, and shows the consistency between the two runs of the same condition.

This resulted in an 8 conditions by 8 conditions symmetric RDM, summarising the representational geometry in the searchlight (e.g. **Figure 5.3**b). The ideal structural representation can be formalised as an 8x8 model RDM, where the distances between conditions are determined by relational structure (**Figure 5.3**a). To test whether the data RDM of a given searchlight was consistent with the model RDM, we calculated the contrast between the means of the data RDM's hypothesised "dissimilar" and "similar" elements (white and black elements in **Figure 5.3**a, respectively). We then used permutation tests to ask whether this contrast was significantly positive across subjects (see section 5.2.55.2.5.1). We repeated this procedure for each searchlight centre on the cortical surface, resulting in a cortical map of p-values.

The only cluster to survive multiple comparisons correction across a cortical hemisphere was located focally in the right entorhinal cortex (**Figure 5.3**b-d top, P=0.045 FWE corrected for cluster mass, cluster-forming threshold P<0.001, peak MNI coordinates: [25,-5,-28]). This effect did not change when we repeated the analysis using model RDMs where same-stimuli or same stimuli set elements were ignored (**Figure 5.4**). The effect was therefore not driven by background colour or low-level plasticity between stimuli that appear in the same block, but rather by a representation of the relational structure between the

stimuli in the task. We note that a cluster in the right temporal-parietal junction (TPJ), where evidence of grid-like coding has been reported (Constantinescu et al., 2016; Doeller et al., 2010), was the strongest and largest amongst the few other clusters surviving the cluster-forming threshold of P<0.001 (but did not survive FWE-correction). To verify that our analysis approach was indeed valid, we tested for stimulus visual identity coding using a similar procedure with an appropriate model RDM (**Figure 5.3**a, bottom). As expected, we observed bilateral effects in visual areas, peaking in the lateral occipital cortex (b-d, bottom, P<0.05 FWE corrected on cluster level, cluster-forming threshold P<0.001, peak MNI coordinates: [44,-74,-4]). Therefore, the relational structure of our task is represented and generalised in EC.

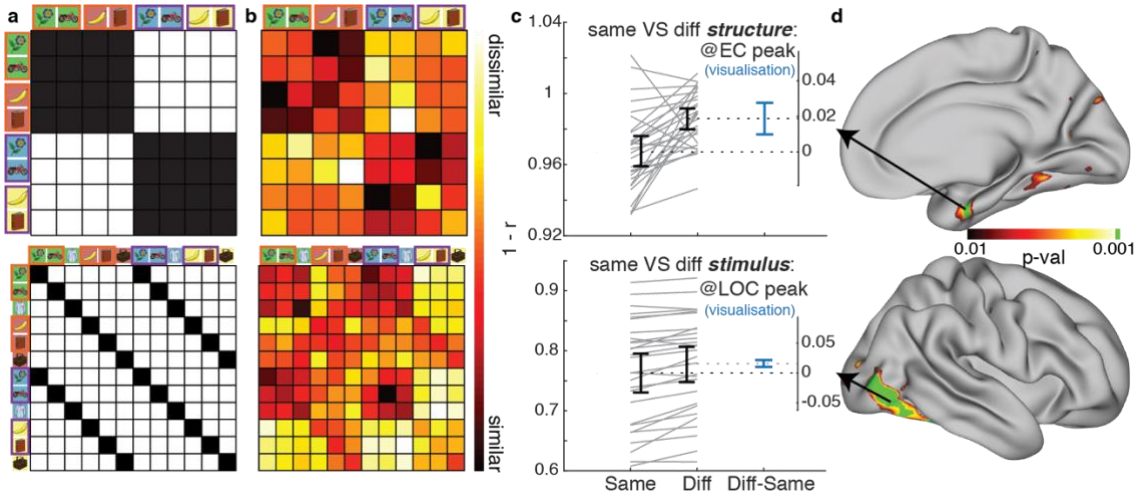


Figure 5.3 The relational structure of the task is represented in the entorhinal cortex. Top: Structure effect, peaking in EC. **Bottom**: stimulus effect, peaking in LOC. **a.** model RDM. Black elements should be similar, white elements should be dissimilar. Pairs of stimuli with orange and purple rectangles around them are +Corr and -Corr, respectively. **b.** Visualisation of the data RDM from peak vertex of the effect, marked with an arrow in **d. c.** Visualisation of the paired mean difference effects between *same* (black RDM elements in **a.**) and *different* (white elements in **a.**) pairs of conditions, from the peak vertex of the effects. Both groups are plotted on the left axes as a slope-graph: each paired set of observations for one subject is connected by a line. The paired mean difference is plotted on a floating axis on the right, aligned to the mean of the *same* group. The mean difference is depicted by a dashed line (consequently aligned to the mean of the *different* group). Error bars indicate the 95% confidence interval obtained by a bootstrap procedure. **d.** Whole surface results, right hemisphere. Clusters surviving FWE correction at a cluster forming threshold of P<0.001 are indicated in green.

It is important to note that there are some inherent differences between the conditions in the task: First, +Corr and -Corr differ in the average reward rate when that average is calculated in short time windows (though across the entire block, as the regressors are defined, the reward rate was controlled and matched between conditions (correlation of

"structure" vector (-1 for -Corr, 1 for +Corr) with a "outcome" vector = 0.04)). Switches of outcomes (regardless of the identity of the stimuli) are more frequent in the -Corr condition than the +Corr condition (correlation of "structure" vector (-1 for -Corr, 1 for +Corr) with a "outcome switch" vector = -0.5, $P<0.001$), though this did not influence subject's tendency to incorrectly make the same choice (compare number of incorrect "stays" between conditions, $P>0.5$). In addition, there is an inherent difference between accept/reject trials as subjects might treat accept trials as riskier. However, there was no evidence that subjects were biased to reject or accept (compare intercepts (across subjects) of a logistic regression with subject's choice as a dependent variable and the STRUCT model outcome probability estimation as the independent variable to 0, $P>0.5$).

5.4 **Discussion**

The ability to appropriately generalise previously acquired knowledge to novel situations is a hallmark of human intelligence. A possible neural solution to this problem is to devote pools of neurons to represent the relations between entities in the environment explicitly, in a manner that is divorced from the entities themselves. Such an explicit representation can generalise to novel situations with the same relational structure. In the previous chapter we presented ideas suggesting that grid cells, originally recorded in the entorhinal cortex, can be thought of as such a representation of the relations between different locations in physical space. Here we use fMRI to show that the human EC explicitly represents relations between reward statistics of stimuli in a reinforcement learning task.

Our experiment can be viewed as a set of non-spatial "remapping" experiments. The horizontal arms in **Figure 5.1**d (same task structure, different stimuli) are analogous to the classic spatial sensory remapping experiments in rodents (Barry et al., 2012; Fyhn et al., 2007), where an animal is moved between different sensory examples of the same task structure (namely free foraging in a Euclidian 2D space). In these experiments, entorhinal grid cells maintain (generalise) their covariance structure across environments (i.e. they do not remap, in contrast to hippocampal place cells (Bostock et al., 1991)). Such generalisation was also observed in Macaque OFC neurons, across different sensory examples of an economic decision-making task (Xie and Padoa-Schioppa, 2016). The vertical arms in **Figure 5.1**d (same stimuli, different task structure) can be viewed as "task remapping" experiments. In the spatial case of such experiments, where animals are required to perform different tasks

in the same sensory environment, recent evidence suggests the grid code changes (remaps) across tasks (Boccara et al., 2019; Butler et al., 2019). Our EC effect (**Figure 5.3**) mirrors these results in a non-spatial RL task in humans. Future work might address whether the exact same neuronal population underlies generalisation in both spatial and non-spatial tasks.

The relational structure between the stimuli is just one aspect of the task. As this is an RL task, we can expect find a number of previously reported RL neural signals. In the next chapter, we use data from the same experiment to explore whether these signals also depend on the task structure.

5.5 **Supplementary information**

Day	Sess	Stimuli Sets per block (each number is a single stimuli set - a unique triplet of images. Triplets sampled independently for each subj.)	Structure Per block (correlation between outcome probabilities of two stimuli, order counter-balanced across subjects)	Stimuli background colour (Blue, Green, Red, Yellow, counter-balanced across subjects)	Trials per block (same for all subjects)	Feedback (same for all subjects)
1	1	5,5,5,5,5,5,5,5	-,-,-,-,-,-,-,-	None	60	Full
1	2	6,6,6,6,6,6,6,6	-,-,-,-,-,-,-,-	None	60	Full
2	1	7,7,7,7,7,7,7,7	+,+,+,+,+,+,+,+	None	60	Full
2	2	8,8,8,8,8,8,8,8	+,+,+,+,+,+,+,+	None	60	Full
3	1	1,1,1,1,1,1,1,1	-,+,-,+,-,+,-,+	B,G,B,G,B,G,B,G	60	No feedback on rejections in last 15 trials

3	2	2,2,2,2,2,2,2,2	+,-,+,-,+,-,+,-	R,Y,R,Y,R,Y,R,Y	60	No feedback on rejections in last 15 trials
4	1	1,1,2,2,1,1,2,2	-,+,-,+,+,-,+,-	B,G,Y,R,G,B,R,Y	60	No feedback on rejections in last 15 trials
Pre-scan	1	2,2,1,1	-,+,-,+	Y,R,B,G	60	Full
scan	1	1,2,1,2,1,1,2,2	-,+,+,-,-,+,-,+	B,R,G,Y,B,G,Y,R	30	Full

Table 5.1 Training schedule for an example subject.

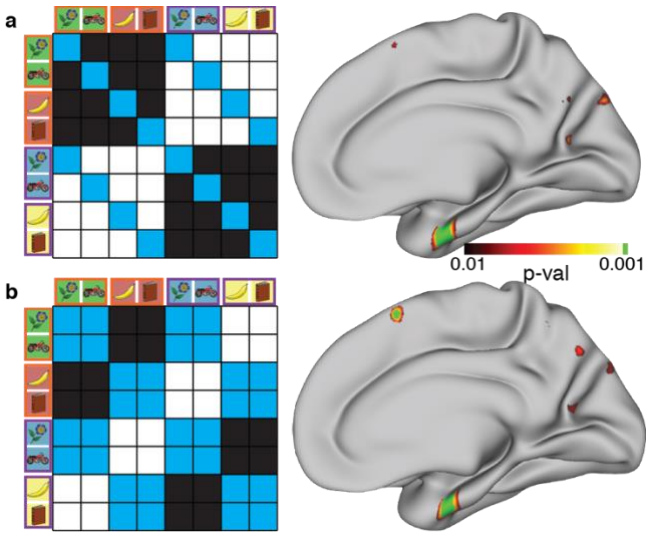


Figure 5.4 Control contrasts for the EC relational structure effect (Figure 5.3, top). Left: model RDMs. Black elements should be similar, white elements should be dissimilar, blue elements are ignored. **Right**: whole-surface results (right hemisphere). colour map is the same as in Figure 2. **a.** Structure effect, ignoring all elements of pairs of same-stimulus conditions. **b.** Structure effect, ignoring all elements of pairs of same stimuli set conditions.

6 STRUCTURE-SENSITIVE REINFORCEMENT LEARNING NEURAL SIGNALS IN THE VENTRAL PREFRONTAL CORTEX

In the previous chapter we introduced a "task-remapping" paradigm in a reinforcement learning setting. We identified representations of the relational structure between task elements – representations we suggest are domain-general. In this chapter, using the same data, we utilise the rich computational theory of reinforcement learning to ask whether neural signals specific to the RL domain are also sensitive to the relational structure of the task. In the first half of the chapter we show indirect and direct evidence for the coding of a latent structural variable. We replicate previous findings (Hampton et al., 2006) *showing that the reward network (particularly vmPFC and HPC) use this variable when encoding value and prediction error signals. We then show, using two independent representational analyses, that the ventral PFC and particularly OFC directly represent this latent variable. These results are compatible with recent work suggesting the OFC represents the current location within the task state-space* (Schuck et al., 2016; Wilson et al., 2014). *In the second half of the chapter I explain the chapter's most exciting result: a dependence of the strong prediction error signals in vmPFC and ventral striatum (vStr) on the relational structure. To date, these signals have been thought of as unitary signals that play the same role in all learning tasks. Our data suggests these signals are maintained in different neural populations (or subspaces) in different tasks, rendering them a part of the brain's mechanism for inference.*

6.1 Introduction

Reinforcement learning (RL) is accredited to some of the most elegant unions of theoretical and experimental work in the history of neuroscience. At the core of the RL framework is the concept of a "state space", comprised of "states" - abstract representations

summarising the external and internal information that is relevant for behaviour, and the (possibly action-dependent) transitions between them (Sutton and Barto, 1998). Consequently, learning to represent a task correctly is a crucial part of learning to solve a task. While classic RL algorithms assume a correct state representation for the algorithms to act on, constructing these can be a hard and resource intensive feat (Gershman et al., 2015; Gershman and Niv, 2010; Niv, 2019). Recent suggestions have identified the ventral prefrontal cortex as a crucial area for the representation of the state the agent/animal is currently in, particularly when the information required to define the correct task-relevant state is latent (Schuck et al., 2016; Wilson et al., 2014).

Once a task is represented with a useful state space, surprisingly simple RL algorithms can, many times, solve it efficiently. For example, tasks that were historically treated as complex model-based tasks, like probabilistic reversal learning (Hampton et al., 2006) or the 2-step task (Daw et al., 2011), can (under the right parameters) be solved with simple model-free algorithms given the right state representation (Akam et al., 2015; Wilson et al., 2014). In environments with a latent structure, neural signals of algorithmic variables operating on the correct state representations (that know about structure) should follow computational models that use the structure (Hampton et al., 2006). The correlated bandits task presented in the previous chapter, which is very similar to a reversal learning task, falls into this category (as the modelling and behavioural results in section 5.3.1 show): with a correct representation of the task (representing the related stimuli A and B as a single latent variable, i.e. $H_{AB} \in \{1, -1\}$), our STRUCT model reduces to a simple Rescorla-Wagner model (Rescorla and Wagner, 1972).

RL algorithms rely on two important (and very related) variables: they estimate the *values* of states (or actions, or state-action pairs), based on *prediction errors (PEs)*. While model-free algorithms only employ a reward PE, model-based algorithms also use state or transition PEs, indicating how observations differ from the current model of the environment (Gläscher et al., 2010).

In the correlated bandits task presented in the previous chapter (section 5.2.2), subjects only experience reward prediction errors[1], as in every trial subjects learn whether they made a correct or incorrect choice. Neural reward PE signals, originating in dopaminergic neurons (Schultz et al., 1997), have been observed in fMRI in a network of reward-related regions, peaking in the ventral striatum (Daw et al., 2011; Delgado et al., 2000; Dupret et al., 2013; Hare et al., 2008; Knutson et al., 2007, 2000; Lohrenz et al., 2007; Peters and Büchel, 2009; Preuschoff et al., 2006) and mPFC (Daw et al., 2011; Hampton et al., 2006; Ramnani et al., 2004; Wimmer et al., 2012). Throughout more than 20 years in the spotlight of the neuroscientific stage, this PE signal was thought to be unitary – the same scalar reward PE signal, regardless of the task at hand or its structure.

Interestingly, much of the network coding for these algorithmic RL variables, and in particular the mPFC, was also highlighted in the previous chapter as an area where in humans grid-like coding was also demonstrated (fMRI: Bao et al., 2019; Constantinescu et al., 2016; Doeller et al., 2010; Unit recordings: Jacobs et al., 2013). If these variables have different consequences in different task structures, they may also conform to the representational predictions (Wang et al., 2018) originating from grid cell activity in remapping experiments (section 5.4): Instead of a unitary representation of PE, these regions might generalise PE representations across different problems with the same task structure, but have different representations across problems with different structures.

6.2 **Methods**

The task, subjects, behavioural modelling and results, and fMRI acquisition and preprocessing details were presented in section 5.2.

[1] As explained in section 5.2.3, the reward PE discussed here is actually the "correctness" prediction error – where a correct choice of the subject is considered rewarding – even if this choice was to correctly reject a trial with a "bad" outcome, thus avoiding monetary loss (but resulting in no monetary reward gained).

6.2.1 Univariate analyses

Due to incompatibility of FSL with the MATLAB RSA toolbox (Nili et al., 2014) used in subsequent analyses, we estimated all first-level GLMs and univariate group-level analyses using SPM12 (Wellcome Trust Centre for Neuroimaging, http://www.fil.ion.ucl.ac.uk/spm). For univariate analyses, contrasts of parameter estimates were smoothed with a kernel of 5mm FWHM before performing group level statistics.

In the following descriptions, [A, B, C] refer to the three stimuli presented in a particular block, where the outcome probabilities associated with A & B was correlated (either positively or negatively, depending on the block).

To test whether chosen action value signals in the reward network were consistent with a model that used the relational structure (section 5.2.3), we constructed a GLM where chosen action value estimates from both STRUCT and NAÏVE models were pitted against each other (**Figure 6.1**a-b), without orthogonalizing them. GLM2 included the following regressors per each block: two main effect regressors of all related stimuli trials ([AB]): one modelling the times of stimulus presentation and the other modelling outcome times. Two parametric regressors were locked to stimulus presentation times of [AB]: the value of the chosen option from the STRUCT and the NAÏVE models. In addition, the GLM included several other regressors: C trials at stimulus presentation; C trials at outcome; 2 regressors modelling button presses across all stimuli: one modelling all "accept" trials and one modelling all "reject" trials; 6 motion parameters as nuisance regressors; bias term modelling the mean activity in each block. **Figure 6.1**a shows the results of the contrast [STRUCT chosen option value] > [baseline]. **Figure 6.1**b shows the results of the contrast [STRUCT chosen option value] > [NAIVE chosen option value].

To test whether prediction error signals in the reward network were consistent with a model that used the relational structure, we constructed a GLM where "correctness" prediction error estimates from both STRUCT and NAÏVE models were pitted against each other (**Figure 6.1**c). GLM3 included the following regressors per each block: two main effect regressors of all related stimuli trials ([AB]): one modelling the times of stimulus presentation and one modelling outcome times. Two parametric regressors were locked to outcome presentation times of [AB]: the "correctness" prediction error estimates from the STRUCT and the NAÏVE models. In addition, the GLM included several other regressors: C trials at stimulus presentation; C trials at outcome; 2 regressors modelling button presses across all stimuli: one modelling all "accept" trials and one modelling all "reject" trials; 6 motion

parameters as nuisance regressors; bias term modelling the mean activity in each block. **Figure 6.1**c shows the results of the contrast [STRUCT chosen option value] > [baseline].

To test for a representation of a behaviourally-relevant, latent AB variable using repetition suppression/enhancement (section 6.3.1.2.2), we constructed a GLM (GLM4) with 9 regressors modelling stimuli presentations times (3 current trial stimuli x 3 preceding trial stimuli): [A(O) => A(S)], [A(O) => B(S)], [A(O) => C(S)], [B(O) => A(S)], [B(O) => B(S)], [B(O) => C(S)], [C(O) => A(S)], [C(O) => B(S)], [C(O) => C(S)]. The notation [A(O) => C(S)], for example, denotes a regressor locked to stimulus presentation of C trials that were preceded by an A trial. We also included several other regressors: all trials at outcome; 2 regressors modelling button presses across all stimuli: one modelling all "accept" trials and one modelling all "reject" trials; 6 motion parameters as nuisance regressors; bias term modelling the mean activity in each block. We divided the 9 trial types into three groups: same ([A(O) => A(S)], [B(O) => B(S)], [C(O) => C(S)]); related ([A(O) => B(S)] , [B(O) => A(S)]); control ([[A(O) => C(S)] , [B(O) => C(S)] , [C(O) => A(S)] , [C(O) => B(S)]). We then calculated the contrast [control] – [related], i.e.: [[A(O) => C(S)] + [B(O) => C(S)] + [C(O) => A(S)] + [C(O) => B(S)]] – 2*[[A(O) => B(S)] + [B(O) => A(S)]]. A positive effect of this contrast is a latent variable suppression effect, while a negative effect (as we observed, **Figure 6.3**a) is an enhancement effect.

To test whether subjects found control trials more difficult than related trials (an indicator task-switching), we examined the reaction times for same, related and control trials by taking the logarithms of the reaction times, Z-scoring them within each block, and then performing a two sided t-test across subjects (**Figure 6.3**b).

See below for a description of the univariate STRUCT model prediction error analysis (without competing with NAÏVE model), the results of which are shown in the insets of **Figure 6.4**b and **Figure 6.4**c.

6.2.2 RSA analyses

General details about all RSA analyses, including the ones in this chapter, are described in section 5.2.5.1.

6.2.2.1 Testing for a latent variable representation

To test for a representation of the latent AB variable (section 6.3.1.2.1 & **Figure 6.2**), we once again used the regression coefficients from GLM1 (section 5.2.5.3). This time, however,

we used the regression coefficients from all three stimuli in each block at both stimulus presentation (S) and outcome presentation (O) times. This was enabled by the long (mean ~4.5 seconds, jittered) S-O delay period. As we were testing for a within-block effect, here we used a different distance metric: a simple correlation distance, calculated within each block. Thus, we obtained three 3x3 data RDMs per block, one for each pair of experimental periods (S-S, O-O, S-O), where the 3 conditions in each RDM correspond to the 3 stimuli presented in the block. Note that each element in the S-O RDM is an average of two correlations, e.g. $d_{A,B} = 1 - mean(corr(A_O, B_S), corr(A_S, B_O))$. We then averaged the RDMs across all blocks (collapsing over blocks of different relational structures). To obtain a single measure of the latent variable representation and to increase statistical power, we averaged the S-S, O-O and S-O RDMs. Next, we defined a hypothesis-driven contrast between RDM elements: unrelated (AC, BC) elements should be more dissimilar to each other than related (AB) elements. We once again used permutation tests for statistical testing and FWE correction (see below for more details).

6.2.2.2 Testing for a prediction error x structure interaction

To test whether standard prediction error signals depended on the relational structure (prediction error x structure interaction analysis, section 6.3.2, **Figure 6.4** and **Figure 6.6**), we first wanted to replicate the previously described ventral striatum (vStr) and vmPFC univariate prediction error signals (Daw et al., 2011; Delgado et al., 2000; Hampton et al., 2006; O'Doherty et al., 2003; Ramnani et al., 2004). We conducted a GLM (GLM5) which included the following regressors per each block: 2 main effect regressors ([AB],[C]) modelling outcome times; one parametric regressor of the "correctness" prediction error from the STRUCT model, locked to the time of [AB] outcome presentation; 2 main effect regressors ([AB],[C]) modelling stimulus presentation times; 2 regressors modelling button presses across all stimuli: one modelling all "accept" trials and one modelling all "reject" trials; 6 motion parameters as nuisance regressors; bias term modelling the mean activity in each block. For each subject, we calculated the contrast [STRUCT AB "correctness" prediction error] > [baseline], smoothed the contrast image using a 5mm FWHM kernel, and obtained group-level results using SPM (**Figure 6.4**b-c, insets). As expected, the two strongest peaks were in vStr and vmPFC. We then used these peaks as (unbiased) ROIs for the multivariate prediction error x structure interaction analysis. To this end, in addition to the surface-based cortical searchlight procedure described above, we also defined 100 voxels long volumetric searchlights within an anatomical mask of the vStr (Harvard-Oxford

Subcortical Structure Atlas). Unlike the surface-based searchlight, the analyses for the volumetric searchlight were all performed in MNI152 space. Only the regressor modelling the AB "correctness" prediction error in each block was used in the RSA analyses for **Figure 6.4** and **Figure 6.6**, resulting in 8 patterns – one pattern per block. We again used the cross-run correlation distance (as described in section 5.2.5) to collapse patterns of the same conditions (a particular stimuli set under a particular structure) across runs, this time to construct 4x4 data RDMs – one condition per block type. Again, we defined a hypothesis-driven contrast between RDM elements, where "different structure" elements should be more dissimilar to each other than "same structure" elements. We report the (uncorrected) p-values obtained by permutation tests in the searchlights centred on the vStr and vmPFC peaks of the univariate prediction error effect (**Figure 6.4**). We also present the exploratory, uncorrected results across cortex of this effect (**Figure 6.6**).

Lastly, we tested whether the vmPFC prediction error x structure interaction effect correlated on a subject-by-subject level with the mOFC latent variable representation effect, by extracting the contrasts of different subjects at the peaks of both effects and correlating across subjects (**Figure 6.5**).

6.3 **Results**

6.3.1 The brain represents and uses a behaviourally useful latent state-space

In this section I present both indirect and direct evidence that in our task, the brain performs RL computations using a useful state space, which includes a latent variable enabling learning from A on B and vice versa.

6.3.1.1 Signals in the reward network and the HPC reflect RL algorithmic variables operating on a state-space with the correct latent structure

We first aimed to test whether known neural signals of RL showed evidence of knowledge about the relational structure of the task. We tested this by comparing how well a model that uses relational structure explained neural signals relative to one that does not use structure. In both models, we calculated the value of the chosen action (accept/reject) on each trial of the two related stimuli (A and B). Note that the chosen action value has the same magnitude as the stimulus value tracked by the models but has an opposite sign on rejected trials. The first model was a naïve Rescorla-Wagner model (NAÏVE, cross-terms in equation 2 set to

zero), and the second model utilised the relational structure (STRUCT, cross-terms in equation 2 fit to behaviour). The chosen action value estimates were used to construct one regressor per model at the time of stimulus presentation (GLM2). Estimates from both models were pitted against each other in the same GLM, meaning any variance explained by a particular regressor was unique to that regressor, allowing us to compare the neural signals uniquely explained by each model. The contrasts in this section only included the two related stimuli in each block, as the differences between the models for the control stimulus were negligible.

A network of regions including the medial prefrontal cortex (mPFC), the amygdala (AMG), the anterior hippocampus (HPC) and the EC coded positively for the chosen action value from the STRUCT model, while a network including the anterior cingulate cortex (ACC), insula, angular gyrus and most of the orbitofrontal cortex (OFC) showed negative coding (**Figure 4.1**a & Table S1; Similar results were obtained for the STRUCT > NAÏVE contrast, **Figure 4.1**b).

We then performed the same analysis, this time with the prediction error parametric regressors, locked to the time of outcome presentation (GLM3). Note that the prediction error used in all fMRI analyses is the "correctness" prediction error, defined as the magnitude of the prediction error from the STRUCT model, and a sign that depends on the congruence between the subject's choice and the outcome: positive when the outcome matches the subject's choice (accept-"good" outcome; reject-"bad" outcome), and negative when choice and outcome are incongruent (accept-"bad" outcome; reject-"good" outcome). The STRCT model prediction error estimates correlated with activity in the ventral striatum, HPC and AMG (**Figure 4.1**c). The same regions coded for the STRCT model more than the NAÏVE model (data not shown).

These results largely form a replication of the work in (Hampton et al., 2006). This indicates that RL neural computations performed by the reward network and HPC operate in a state-space that has the correct (latent) relational structure of the task.

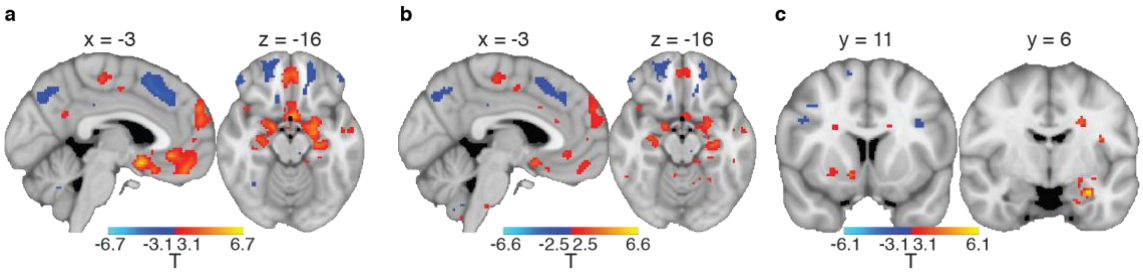


Figure 6.1 Reward network signals reflect RL algorithmic variables operating in a state-space with the correct latent structure. a. regression coefficients for chosen action value estimates from STRUCT model (when competing with NAÏVE model in the same GLM). **b.** Same as **a.** for STRUCT > NAÏVE contrast. **c.** regression coefficients for "correctness" prediction error estimates from STRUCT model (when competing with NAÏVE model in the same GLM).

6.3.1.2 OFC representation of task space respects the relational structure.

In many laboratory tasks, including ours, more than one task representation is possible (Niv, 2019; Wilson et al., 2014). However, to benefit from the relational knowledge, any such representation must differentiate the related stimuli (A,B) from the control stimulus (C) so that AB outcomes can cause learning on both A and B but not C. For example, one possible representation is to maintain a single variable encoding the state of the AB pair (Wilson et al., 2014) and a second variable encoding the state of C.

6.3.1.2.1 RSA

We therefore tested for regions where the representations of AB trials were more similar than AC or BC, again by constructing RDMs. To increase statistical power, we consider representational similarity at stimulus time (S-S), at outcome time (O-O) and between stimulus and outcome (S-O). It is particularly interesting for credit assignment if representations generalise from outcome time to the time when the stimulus was presented. Note that unlike other RSA analyses in this thesis, these RDMs are defined within-blocks - not between blocks - made possible as ABC trials are presented in pseudo-random order. Hence, within each block, we used correlation distances to calculate three 3x3 data RDMs (S-S, O-O, S-O), where the three conditions in each RDM correspond to the three stimuli. We then averaged the RDMs across all 8 blocks, and defined contrasts between unrelated (AC, BC) and related (AB) pairs of stimuli. Finally, we averaged the three contrasts to obtain a single measure of the latent AB variable representation and submitted it to permutation tests as described above.

The strongest effects in cortex (**Figure 6.2**a) were observed in the medial OFC (Brodmann areas (BA) 11,13,14 P=< 0.0002 uncorrected, P = 0.039 FWE cluster-corrected

in frontal-temporal mask) and the right posterior parahippocampal (pPHC) gyrus (P=0.0005 uncorrected, P=0.0445 FWE cluster corrected for whole surface). All three experimental periods RDMs (S-S, O-O, S-O) contributed to the effect (**Figure 6.2**b). See section 6.5.1 for description and discussion of statistical correction.

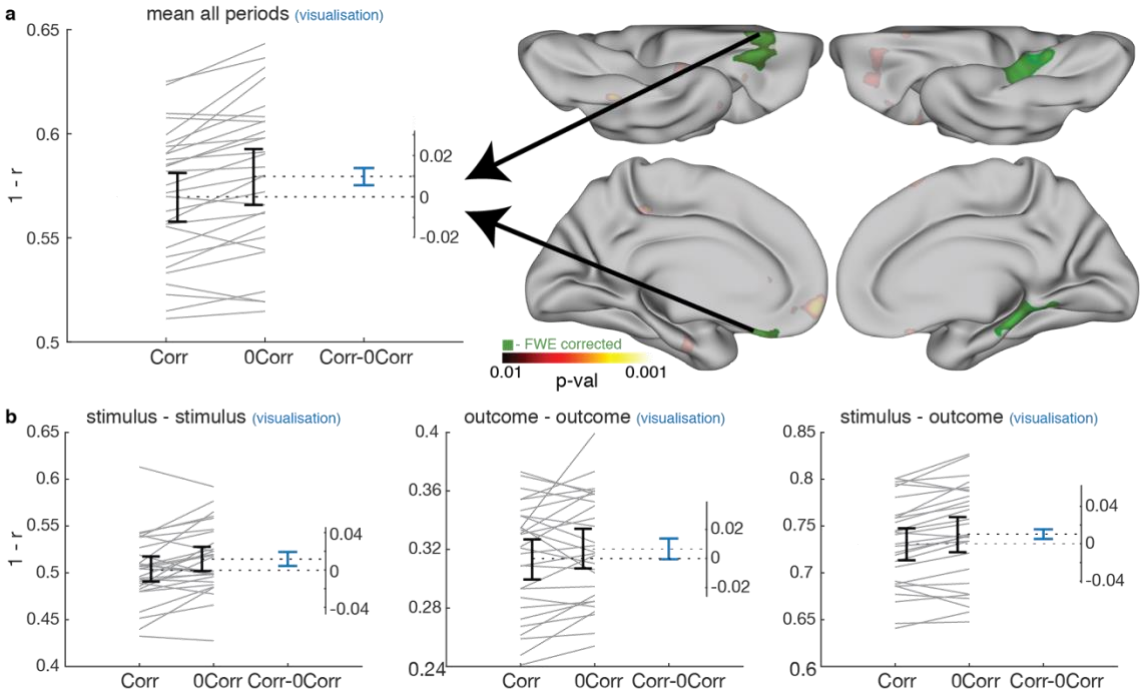


Figure 6.2 OFC represents a latent variable, respecting the task structure. a. Right: Whole-surface results (top/bottom: orbital/medial views) for the [unrelated stimuli – related stimuli] contrast, averaged over all experimental periods (S-S, O-O, S–O). The semi-transparent uncorrected p-value map is overlaid over the two clusters that survived FWE correction, shown in green. Left: Visualisation of the effect from the peak of the OFC cluster, located in the left medial OFC (BA14). **b.** Same as **a. (left),** separated to the different experimental periods (S–S, O-O, S-O). Legends are the same as in **Figure 5.3**.

6.3.1.2.2 Repetition enhancement

The S-O analysis from the previous section (**Figure 6.2**b right) suggests there is an overlap between neuronal populations representing stimulus A during stimulus presentation and stimulus B during outcome time, and vice versa. Such an overlap is particularly useful for credit assignment purposes. We tested directly for such an overlap using the independent analysis method of repetition suppression/enhancement (Barron, Garvert, & Behrens, 2016, GLM4). An overlap between the AB populations should result in a modulation of signal during stimulus B presentation (denoted B(S)) after (in the last trial) the outcome for stimulus A (A(O)) was presented (or vice versa), whereas such a modulation should not be observed

during (for example) stimulus B presentation (B(S)) following previous trial outcome for the control stimulus C (C(O)).

We observed such an effect across the entire orbital surface, peaking in bilateral ventrolateral PFC and mPFC (cold colours **Figure 6.3**). Importantly (and somewhat unexpectedly), the effect was an enhancement effect, rather than a suppression effect: the vPFC signal in AB trials following other AB trials (*related trials*: not including same-stimulus trials, i.e. only including [B(O) => A(S)] and [A(O) => B(S)] trials) was larger than in trials involving the control stimulus (*control trials*: [B(O) => C(S)], [A(O) => C(S)], [C(O) => A(S)], [C(O) => B(S)]). While this effect did not survive multiple comparisons correction and should hence be considered exploratory, it has several features that make it believable: with a low threshold, it closely follows the grey matter of the vPFC and it is very bilateral (cold colours in Figure 6.3a).

While this might be a latent variable effect, plausible interpretation of this effect is as a task-switching related activity decrease. Under this view, A or B trials are part of one task, and C trials are part of another task. Indeed, two indications point towards this interpretation: 1) There was an increase in activity in the dorsal frontoparietal network when switching tasks (control trials compared to related trials), as suggested by the literature (Dove et al., 2000; Kimberg et al., 2000) (hot colours in Figure 6.3a). 2) Reaction times of the control pairs of stimuli were higher than for related pairs, though this difference was not significant (Figure 6.3b). In additions, all of these effects are quite weak and do not survive multiple comparisons correction. Hence, these results, while consistent with previous literature, are exploratory at best.

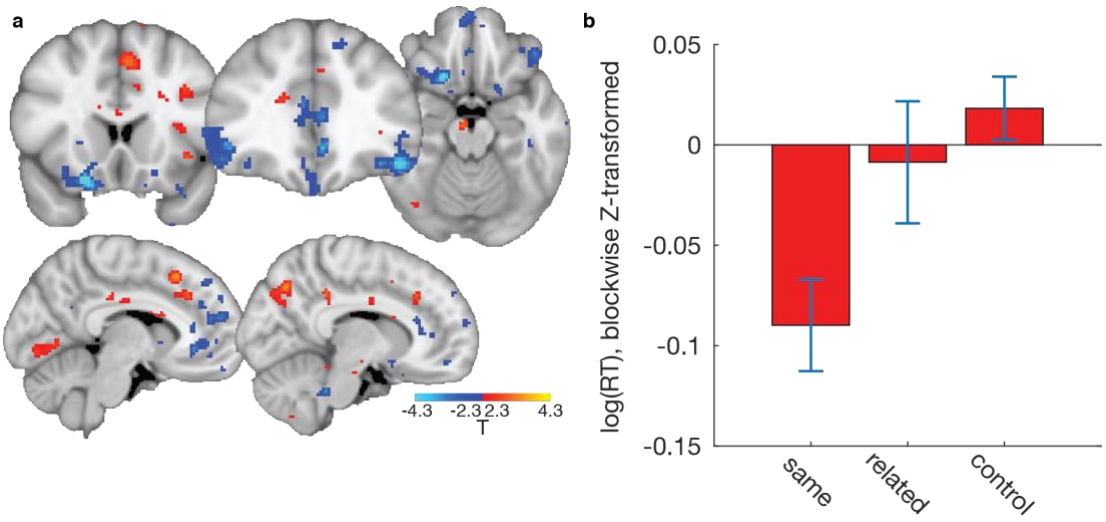


Figure 6.3 Possible latent variable repetition enhancement effects in the OFC and vlPFC. a. Repetition suppression and enhancement effects for the [control – related] contrast. Hot colours are suppression effects, cold colours are enhancement effects. Note the low threshold (T>2.3 uncorrected). **b.** *log*(Reaction times), z-scored within each block, for each of the three types of consecutive stimuli. Error bars show SEM across subjects.

6.3.2 vmPFC and ventral striatum represent the relational structure in learning signals

The vmPFC is central to reward-guided learning (Rushworth et al., 2011). Intriguingly, previous studies suggest grid-like coding can also be found in mPFC (Bao et al., 2019; Constantinescu et al., 2016; Doeller et al., 2010; Jacobs et al., 2013). We hypothesised that vmPFC learning signals of the type typically observed in RL tasks (Daw et al., 2011; Hampton et al., 2006; Ramnani et al., 2004; Wimmer et al., 2012) may be sensitive to the relational structure. That is, because prediction errors have different implications for learning under the two different relational structures, we hypothesized that vmPFC would not simply encode a signal that monotonically increases with prediction error as previously reported. Instead, we hypothesized that the representation of prediction errors across voxels would differ depending on the relational structure.

We first found a strong univariate prediction error signal (GLM5) in a network of regions including vmPFC (**Figure 6.4**b, inset) and the ventral striatum (vStr, **Figure 6.4**c, inset), in line with previous findings (Daw et al., 2011; Hampton et al., 2006; Ramnani et al., 2004; Wimmer et al., 2012).

We next asked whether the multivariate pattern of this prediction error signal depends on the relational structure on a fine-grained scale, in a multivariate analysis. We conducted a searchlight RSA analysis similar to sections 5.2.5 and 5.3.2, with two notable differences: A) the patterns used as inputs to the RSA were not the average responses to the stimuli, but instead the regression coefficients of the prediction errors on the two related stimuli (A and B) in each block. This means the patterns entering this analysis are the local spatial variations in the representation of prediction errors. B) These analyses only have a single measure per block (prediction error coefficients) as opposed to two (separate stimuli) in the previous section (GLM5). The resulting RDMs are therefore 4x4, not 8x8.

Because we are testing the multivariate differences on the (orthogonal) univariate prediction error effect, we could use the peaks of the univariate effect to constrain our regions of interest. The top two univariate peaks were in vmPFC (peak MNI [-4,44,20]) and ventral striatum (vStr, peak MNI [-10,8,-12]. In searchlights centred on both these peaks, the

multivariate prediction error X structure interaction effect was significant (vmPFC P=0.014, **Figure 6.4**a and **Figure 6.4**b; vStr P=0.034, **Figure 6.4**c). We also note that other regions where grid-like coding was previously observed, such as the posterior cingulate cortex (PCC) (Constantinescu et al., 2016; Doeller et al., 2010) showed this interaction effect, as well as other reward-related regions like the ventrolateral PFC (Rushworth et al., 2011). See **Figure 6.6** for a description of these exploratory results.

These results indicate that prediction error signals in vmPFC and vStr (and perhaps vlPFC and PCC) depend on the current relational structure of the task. The critical difference between the two relational structures in our experiment is not how the prediction error should be computed, but rather how it should be used to inform future behaviour – how should 'credit' for the error be assigned (Gershman et al., 2015; Gershman and Niv, 2010; Jocham et al., 2016; Rudebeck et al., 2017; Walton et al., 2010; Wilson et al., 2014). One intriguing possibility is therefore that different representations of prediction errors allow different credit assignment for the same prediction errors in the two relational structures. Factorising these computations from the sensory particularities of the task allows them to be rapidly 'remapped' to new stimuli for rapid learning.

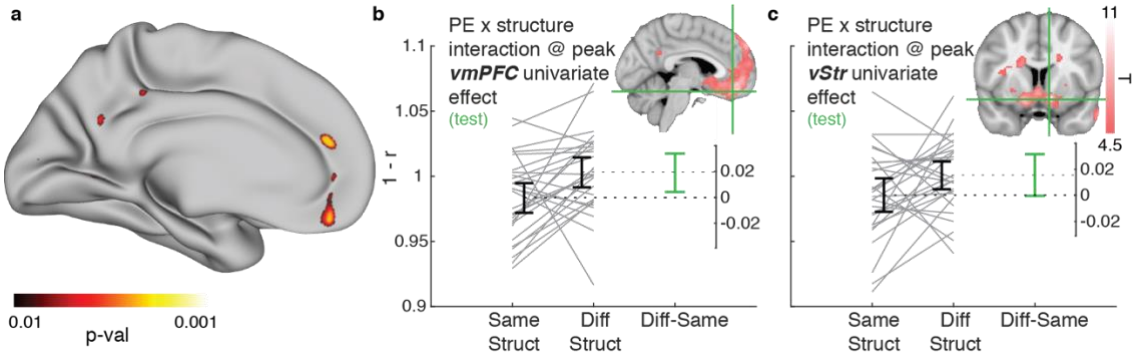


Figure 6.4 Prediction error signals in vmPFC and ventral striatum depend on the current relational structure of the task. a. Visualisation of whole-surface results of the multivariate prediction error x structure interaction effect, medial left hemisphere. **b.** Interaction effect at the vmPFC peak univariate prediction error effect (MNI: [-4,42,-20]). **c.** Interaction effect at the ventral striatum peak univariate prediction error effect (MNI: [-10,8,-12]). Insets in **b.** and **c.** show the univariate prediction error effect. Legend for panels **b.** and **c.** is the same as in **Figure 5.3**.

Intriguingly, the vmPFC prediction error x structure interaction effect at its peak correlated on a subject-by-subject level with the periods-averaged latent variable effect at its spatial peak (mOFC) (Fig. S4, r=0.5, P=0.01, vmPFC MNI [-8,46,-16], mOFC MNI [-6,30,-

26])). Though causal directionality cannot be inferred from this correlation, it is possible that subjects with a vmPFC representation that structurally differentiates credit assignment of errors can utilise it to create a stronger latent variable representation in mOFC – i.e. a better task state-space representation.

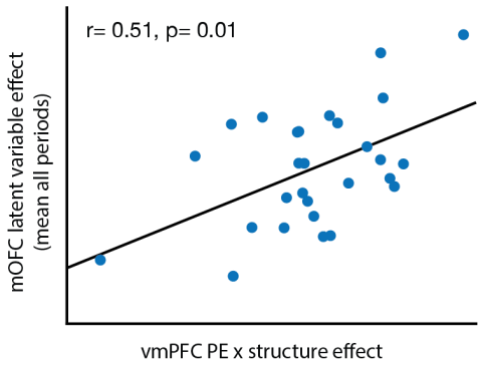


Figure 6.5 Subject-by-subject correlation between the prediction error x structure interaction effect size at the vmPFC peak of the that effect (Figure 3.3a) and the within-block, latent variable effect size at the mOFC (BA14) peak of the effect (Figure 6.2a).

6.4 **Discussion**

In the first half of this chapter, we present evidence that the brain represents and uses a latent variable controlling the correlated outcome probabilities in our task. We show that algorithmic neural signals in the reward network and the HPC are more consistent with a model that uses this structural variable. We then show, using both RSA and repetition enhancement, that the vPFC, and in particular OFC, represent this variable, probably through overlapping representations of the stimuli and outcomes controlled by it. All of the analyses in this part of the chapter were performed within blocks. They did not include comparisons between the two different relational structures in the task (+Corr/-Corr), but rather only comparisons between stimuli which are part of any structure (AB) and the control stimulus (C). In the second half of the chapter we tested whether these signals are also sensitive to the particular relational structure. We suggest a novel interpretation of the heavily studied prediction error signals in vmPFC and vStr, by demonstrating that the fine-grained patterns of the signals vary as a function of the specific relational structure. These two sets of findings are highly related: The precise computations that mediate model-contingent credit assignment and learning depend on the brain's internal representation of the task state-space (Gershman et al., 2015; Gershman and Niv, 2010; Niv, 2019; Schuck et al., 2016;

Walton et al., 2010; Wilson et al., 2014). Strengthening this notion, we finally show that the mOFC latent variable effect and the vmPFC prediction error x structure interaction effect correlate with each other across subjects.

The finding that the reward network coding of chosen action value and predictions errors is consistent with a model that knows about the correlation structure is an almost exact replication of (Hampton et al., 2006). However, our interpretation of the result is slightly different. Hampton et al. present this finding as evidence that the network (and particularly vmPFC, on which they focus in their paper) employs model-based decision making, as it does not only track the simple action-outcome association, but rather it knows about the higher-order structure of the task. Following (Akam et al., 2015; Wilson et al., 2014), we suggest that this task (like other variants of the reversal learning task (Wilson et al., 2014)) can be solved using model-free algorithms, as long as they operate on the correct latent state-space of the task. We then directly show that the crucial part of that state-space – the latent AB variable, is indeed represented in OFC (and pPHCg), compatible with suggestions that this region codes for the current state, in the case that it is latent (Schuck et al., 2016; Wilson et al., 2014). It is important to note that our analysis finds a representation of the latent AB variable, which at any given block can be in one of two states (e.g. in -Corr blocks, the states are [probability of a good outcome on A is high; probability of a good outcome on B is low] and [probability of a good outcome on A is low; probability of a good outcome on B is high]). I elaborate on this difference and my interpretation of our results in the general discussion (section 8.1).

Our latent AB variable repetition suppression/enhancement analysis revealed, somewhat unexpectedly, an enhancement effect in the medial and ventral PFC (**Figure 6.3**a). While the majority of fMRI studies report suppression effects for overlapping representations, including in frontal cortex (e.g. (Barron et al., 2013; Klein-Flügge et al., 2013)), some studies report enhancement (e.g. (de Gardelle et al., 2013; Muller et al., 2013)). It is possible that the enhancement effects we observed in ventral and medial PFC are actually task-switching activity decreases (Rushworth et al., 2002), i.e. switching from the AB task to the C task, and hence these results should be interpreted with care.

It is noteworthy that most subjects could not learn the within-block structure (2 stimuli have a correlated outcome probability, while the third stimulus has an independent outcome probability) without being explicitly told that the structure exists (piloting data – data not shown). This is probably due to the large hypothesis space that is possible with binary

outcomes for on 3 stimuli – particularly when the stimuli are not presented concurrently. This means there can be long periods of time between occurrences of the same or related stimulus, making it difficult for the subject to form and test hypotheses about the task structure. However, the fact that subjects were explicitly informed of the task structure does not undermine the interpretability of our results (though perhaps our effects might have been different or stronger without the explicit information). There is still no explicit information in each trial about the relational structure of the task – subjects must rely on their memory of the task structure, meaning the latent AB variable is indeed latent.

In the previous chapter we suggested a unified framework for the representation of task structure across spatial and abstract tasks. Such a framework might also afford a new way to interpret standard RL neural signals like prediction error. Our vmPFC prediction error x structure interaction effect (section 6.3.2) brings together observations about vmPFC function from several seemingly disparate fields, including RL (Rushworth et al., 2011). Patients with vmPFC lesions show a selective valuation deficit when value comparison should be based on attribute configuration, i.e. when relationships between object elements are important for the valuation (Pelletier and Fellows, 2019). In the memory literature, vmPFC has been strongly implicated in the representation of schemas – abstract structures of previous knowledge, which bear many parallels to the relational structures discussed here (Baldassano et al., 2018; Gilboa and Marlatte, 2017). vmPFC is particularly important when new information is assimilated into an existing schema (Sommer, 2017; Tse et al., 2011), analogous to a prediction error update of the internal model of the task within the current structure. Finally, some of the strongest effects of spatial and nonspatial grid-like coding in fMRI were recorded in the medial PFC (Constantinescu et al., 2016; Doeller et al., 2010; Jacobs et al., 2013). Notably PCC, where we found the strongest prediction error x structure interaction effect in our exploratory analysis (**Figure 6.6**), also exhibits grid-like coding (Constantinescu et al., 2016; Doeller et al., 2010) and has also been strongly implicated in the representations of schemas (Baldassano et al., 2018; Sommer, 2017).

In addition, we show a relationship between the vmPFC structural learning signal and the representation of a behaviourally relevant latent variable in OFC (Gershman et al., 2015; Gershman and Niv, 2010; Schuck et al., 2016; Wilson et al., 2014). Both of these representations are useful for facilitating appropriate model-contingent credit assignment (Jocham et al., 2016; Rudebeck et al., 2017; Walton et al., 2010).

6.5 **Appendix**

6.5.1 A note regarding multiple comparisons corrections in this chapter

Whilst the effects in the latent variable representation analysis (section 6.3.1.2.1) do not survive multiple comparisons correction with the same parameters as the entorhinal effect in section 5.3.2, they do survive with a lower cluster forming threshold of P<0.01 (pPHC gyrus: P=0.0445 whole-surface corrected; mOFC: P = 0.046 small volume corrected in an anatomical mask of the frontal and temporal lobes based on the MNI Structural Atlas in FSL), indicating that the effects are weaker but larger than the EC effect in section 5.3.2. This is perhaps not surprising as OFC and pPHC are larger than EC and therefore can support larger, more diffuse clusters. However, in the absence of pre-registration, we note that the latent variable FWE results in section 6.3.1.2.1, whilst strong, involved minor parameter search and therefore may not control precisely for family-wise errors.

Peak MNI cords	Brain region(s) of cluster	Peak t-score (df=27)	p-value (FWE-corrected on cluster level)
-32,-16,-20	L anterior hippocampus + amygdala + entorhinal cortex	6.71	P<0.001
-4,10,-12	Posterior mPFC (subcallosal cortex)	6.45	P=0.001
-6,58,36	Dorsal mPFC	4.56	P<0.001
28,-44,64	L Somatosensory cortex	5.56	P=0.006
0,34,-8	vmPFC	5.53	P<0.001
-32,-20,2	Insula/parietal operculum/white matter	5.07	P<0.001

26,4,-22	R hippocampus + amygdala + entorhinal cortex	5	P=0.003
52,-32,20	R parietal operculum	4.88	P<0.001
8,34,42	ACC	6.92	P<0.001
-42,16,6	L Insula	6.68	P<0.001
46,-52,34	R Angular gyrus	6.03	P<0.001
40,44,-2	OFC	5.69	P=0.002

Table 6.1 Effect of chosen action value of STRUCT model, when competing with NAÏVE model in the same GLM (GLM2, Figure 6.1a-b). The contrast is [STRUCT chosen action value] > Baseline. All clusters with a FEW-corrected P-value < 0.05 are reported. Note that negative effects (with a negative t-score) are also reported.

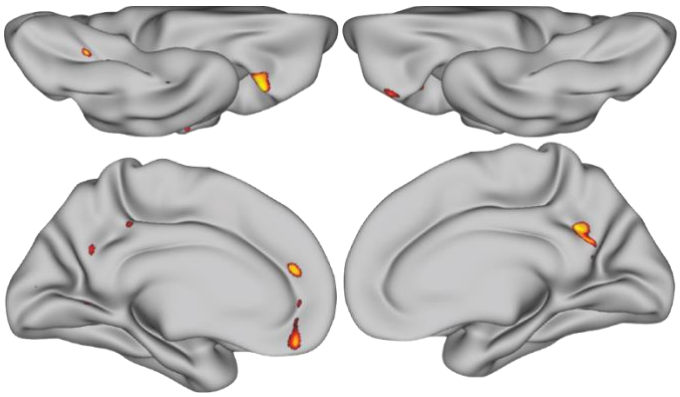


Figure 6.6 Exploratory results of the prediction error x structure interaction effect. Visualisation of whole-surface results of the multivariate prediction error X structure interaction effect (section 6.3.2; top/bottom: orbital/medial views, effects do not survive FWE-correction across a cortical hemisphere). The strongest effects in cortex where observed in PCC RH (right hemisphere): P<0.001 uncorrected), vlPFC (LH: P < 0.001 uncorrected, P=0.002 at the peak of the univariate prediction error vlPFC effect) and vmPFC (LH: P=0.002 uncorrected, **Figure 6.4** a). In all of these regions, the effects, albeit weaker, were also observed bilaterally (PCC LH: P = 0.005 uncorrected; VLPFC RH: P=0.005 uncorrected); vmPFC RH: P=0.01 uncorrected). This network of regions is of particular interest for this effect, as it brings together several seemingly disparate literatures. The vmPFC and vlPFC are central to reward-guided learning (Rushworth et al., 2011). The vmPFC (Gilboa and Marlatte, 2017; Sommer, 2017; Tse et al., 2011), and PCC (Sommer, 2017) have been

strongly implicated in the representation of knowledge schemas. In addition, grid-like coding was reported in both vmPFC and PCC (Constantinescu et al., 2016; Doeller et al., 2010).

7 INFERRING DYNAMIC RELATIONAL STRUCTURES UNDER UNCERTAINTY

In the chapters 5 and 6 I presented evidence for abstracted and generalised representations of task structures in the brain. Importantly, the relational structure of the task was constant within each block and was explicitly signalled to the subjects. In the real world, structures need to be inferred from ongoing observations and their statistical regularities. In this chapter, I present a 2-stimuli variant of the reversal learning task from chapters 5 and 6, where the correlation structure between the stimuli's associated outcomes also changes. In addition, I present a Bayesian model which infers the current structure and the outcome probabilities in an optimal manner. Preliminary behavioural results indicate that subjects infer and use the structure of the task in a close-to-optimal fashion. The close

fit between the model and subject's behaviour presents exciting future opportunities for a model-based fMRI experiment.

7.1 **Introduction**

Humans and animals can rapidly learn new problems and adjust their behaviour when the task environment changes. One important reason that enables them to achieve this feat is their ability to use the high degree of structure in natural tasks (Gershman and Niv, 2010; Kemp and Tenenbaum, 2009). Indeed, usually the relationships between task elements are fairly stable even if the identity of the elements in these relationships changes, enabling fast inferences and generalisations to novel situations with a known relational structure. However, in some tasks the relational structure might change and needs to be inferred. Moreover, knowing the uncertainty in this inference process would be useful, as the extra knowledge usually afforded by knowledge of the task structure should be weighted down if there is high uncertainty about it, and weighted up if it is more precise than other sources of information.

(Boccara et al., 2019) developed an elegant example of such a "dynamic structure" task in physical space (reviewed in section 3.1.2; Figure 3.2c,d). Rats were trained to daily learn three new reward locations on a cheeseboard maze. This can be thought of as a task with hierarchical structure: a high-level structure, which is stable throughout the different days of the task: "there are exactly 3 important locations"; and a lower-level structure indicating the relationships between locations in each day, when reward locations are constant. The low-level structure is dynamic: it changes from day to day, requiring inference on the structure itself. (Boccara et al., 2019) show that animals develop separate and discrete entorhinal grid maps (suggested to encode the structure of tasks (chapter 5)) for each new reward configuration. Moreover, during the learning of the new reward locations, the grid cell population code discretely flips between representations of the old and new structure in an inference-like manner, culminating in a representation that corresponds to the new reward configuration (Figure 3.2c,d).

In chapters 5 and 6 I presented evidence that the same brain regions that encode the structure 2D physical space also encode the structure of a non-spatial reinforcement learning task. I hypothesised that the same mechanisms underlying the inference of the task structure

in physical space (as in (Boccara et al., 2019)) will also underlie inference on relational structures in abstract RL tasks – where such "structure volatility" might be more common than physical space.

It is well known that the brain tracks behaviourally-relevant higher-order statistical properties of RL tasks. For example, (Behrens et al., 2007) showed that in a one-armed bandit task, humans use the volatility of the reward environment to determine the weight that outcomes associated with past decisions should get in determining future decisions. This was shown by comparing subject's behaviour to that of a Bayesian learner which optimally uses the history of outcomes to track the outcome probability.

In this chapter, I present a substantial extension of that model to a correlated bandit task similar to the one presented in chapters 5 and 6. Crucially, here the correlation structure between the outcome probabilities was dynamic and had to be inferred by the subjects on-the-fly, similar to the inference process performed by rodents on the structure of the spatial reward-location tasks in (Boccara et al., 2019). The model tracks the outcome probabilities and the volatility as in (Behrens et al., 2007), but also includes a "structure variable" which determines how to update the outcome estimates of one stimulus following an outcome on the other. In addition, I present (very) preliminary pilot data showing that the Bayesian learner with the true generative model of the task explains the behaviour of human subjects, suggesting they can solve this task and use the different high-order statistics to make choices in an (almost) optimal manner.

7.2 **Methods**

7.2.1 Task

I adapted the task presented in section 5.2.2 to a scenario where the correlation structure between the outcome probabilities associated with stimuli A and B can dynamically change (flip between +Corr and -Corr). This version did not include the control cue C and can be thought of as a probabilistic reversal learning paradigm with 2 stimuli, only with correlation structure reversals as well as outcome probabilities reversals (with the latter being more frequent).

In each trial, subjects were presented with one of two possible objects (A and B). The object was associated with one of two "cards" (arbitrary outcomes) – *green* or *blue*, according

to a predetermined outcome schedule. Subjects had to indicate, using the arrow keys (left – *green*, right – *blue*), which card they believed the object was currently associated with (**Figure 7.1** left). They then received feedback on their choice, informing them whether it was correct or incorrect, and hence at the end of each trial knew what the correct answer was (**Figure 7.1** right).

Crucially, the outcome schedule was determined such that the stimuli's outcome probabilities at any given time were correlated – either positively or negatively. In other words, during +Corr periods, when (for example) A was likely to lead to a *green* outcome, B was also likely to lead to a *green* outcome. In -Corr periods, when A was likely to lead to a *green* outcome, B was likely to lead to a *blue* outcome. Thus, participants could always learn from one stimulus on the other, but in order to know in which direction they had to know the current structure. Each subject completed 400 trials (200 of each stimulus, pseudo-random order) with a schedule sampled from the following distribution: each schedule had 5 switches of the correlation structure, resulting in 3 periods of -Corr periods and 3 periods of +Corr periods. The outcome probabilities of the two cues flipped between 0.95 and 0.05, with the intervals between outcome probability flips sampled from a normal distribution with a mean of 6 trials, standard deviation of 3 and truncated between 3 and 9 inclusive. Note that the difference between an outcome probability switch and a structural switch is that in the former the probabilities associated with *both* stimuli flip (thus preserving the correlation structure), while in latter only the probability associated with one of the stimuli flips (thus flipping the correlation structure). See **Figure 7.3**a for an example schedule.

This task setup is very similar to the one presented in section 5.2.2, with several important differences: 1) The correlation structure changes dynamically. 2) There are only two stimuli rather than three. 3) Subjects do not try to learn associations between the stimuli and reward, but rather associations between stimuli and arbitrary outcomes (green/blue cards). This means subjects do not accept or reject each trial, but rather indicate the outcome they believe is currently associated with the stimulus. The accept/reject and green/blue setups are completely analogous – the change was made for reasons of clarity and simplicity.

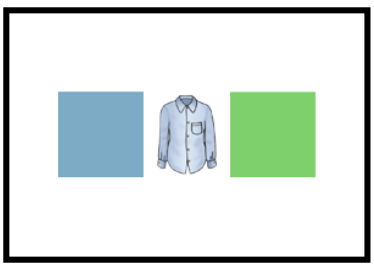




Figure 7.1 Trial structure. In each trial subjects were presented with one of two objects (**left**; here the object is a shirt) and had to use the arrow keys to indicate whether they predict the object is currently associated with the blue or green card. Subjects then received feedback indicating whether their choice was correct or incorrect (**right**; sack of gold indicates correct choice), as well as feedback on their current score (indicated by the number of coins).

7.2.2 Formal description of Bayesian Learner

In the following section, I describe a Bayesian approach for tracking two outcome probabilities given past history of stochastic trial outcomes, in an environment where these probabilities might change and may be correlated, with a possibly changing correlation structure. I follow the work of (Behrens et al., 2007) where an optimal Bayesian learner which tracks the outcome probability of a single, possibly changing probability was presented, and extend it to the scenario of two, possibly correlated, outcome probabilities, with a correlation structure that can possibly change too.

7.2.2.1 Generative model

At trial t a new outcome is generated, together with a context (stimulus) signal, $S \in A, B$. I denote outcomes from context A by y_t^A and outcomes from context B by y_t^B. In our experiment, A and B are different stimuli presented before the outcome, representing two different stochastic processes. $y_t^A = 1$ ("green outcome") with probability g_t^A, and $\mathrm{y_t^A} = 0$ ("blue outcome") with probability $1 - g_t^A$. The same holds for y_t^B and g_t^B. Note that any specific trial can only be either an *A* or a *B* trial. I denote by $\mathrm{y_t}$ and g_t the observed outcome and outcome probability in trial t respectively, regardless of the context of the specific trial. Assuming that on trial t the outcomes are drawn independently from a Bernoulli distribution parameterised by $\mathrm{g_t}$, we can write the likelihood functions:

$$P(y_t^A | g_t^A, g_t^B, S_t = A) = \begin{cases} g_t^A & if\ y_t^A = 1 \\ 1 - g_t^A & if\ \ y_t^A = 0 \end{cases}$$

$$P(y_t^B | g_t^A, g_t^B, S_t = B) = \begin{cases} g_t^B & if\ y_t^B = 1 \\ 1 - g_t^B & if\ \ y_t^B = 0 \end{cases}$$

Equation 7.1

The goal of the subject is to estimate the outcome probability $g_{t+1}^{S_{t+1}}$ given all past observations $y_{1\ldots t}$ and the context S_{t+1}, in order to make an optimal choice on trial $t + 1$. As highlighted by (Behrens et al., 2007), if g is changing in time, it is beneficial for the subject

to understand how it is changing, so that the volatility of the environment influences the discounting of information with passing trials. Moreover, if g^A and g^B are correlated, it is beneficial for the subject to understand the nature of the correlation, so that they can use the information from past A trials to choose optimally in the next B trial and vice versa. For example, if g^A and g^B are perfectly (either positively or negatively) correlated, the outcomes in A trials are as informative as the outcomes in B trials for an optimal choice in the next (either A or B) trial. If g^A and g^B are uncorrelated (independent), there is no information in any A trial outcome that is relevant for choice on any B trial, etc.

I assume that the problem can be solved in a Markovian fashion, such that the environment can be well represented without needing to store the entire history of estimated outcome probabilities. Under this assumption, it is sufficient to know only the last estimated outcome probabilities of A and B, and the correlation relationship between them, in order to make a choice on the next trial. **Figure 7.2** shows the generative model for the correlated Markovian probability-tracking problem. The outcome probabilities of both contexts (stimuli) are being represented simultaneously: $P(g_{t+1}^{same}|g_t^{same})$ and $P\left(g_{t+1}^{other}\middle|g_t^{same}, g_t^{other}, M_t\right)$, where the context probability depends on the context of the current outcome: $g_t^{same} = g_t^A$ and $g_t^{other} = g_t^B$ if $S_{t+1} = A$, and analogously for g_{t+1}^{same} and for $S_{t+1} = B$. M_t is the correlation between g_t^A and g_t^B at trial t.

$P(g_{t+1}^{same}|g_t^{same})$ is centred on g_t^{same}, reflecting the prior assumption that the outcome probability in the next trial of the same context depends on its outcome probability in the current trial. In correlated environments, the distribution on g_{t+1}^{other}, though also centred on g_t^{other}, is also influenced by g_t^{same}, and the strength of that influence is governed by M_t. In the following, I formally present the details of the forward model for $S_{t+1} = A$. The case where $S_{t+1} = B$ is completely analogous. To model possible changes in the outcome probabilities, we used the "JUMP" volatility model from (Behrens et al., 2007), which is optimal for environments where the outcome probability is constant for a period of time, and then jumps to a new unknown value. I developed a novel form of this model, which can take into account the correlations between two outcome probabilities.

Note that g_{t+1}^A and both g_t^B and M_t are conditionally independent given $S_{t+1} = A$ and g_t^A, so $P(g_{t+1}^A|g_t^A, g_t^B, M_t, S_{t+1} = A) = P(g_{t+1}^A|g_t^A, S_{t+1} = A)$. Therefore, when $S_{t+1} = A$ this conditional distribution reduces to the single-context case, as in (Behrens et al., 2007).

As explained in their "JUMP" volatility model, we first need to introduce a latent variable L_{t+1} which takes the value 1 when a jump in outcome probability occurs in trial $t+1$ and 0 when it is stable. The conditional probability distribution on g^A_{t+1} is:

$$P(g^A_{t+1}|g^A_t, L_{t+1}, S_{t+1} = A) = \begin{cases} \delta(g^A_{t+1} - g^A_t) & if\ L_{t+1} = 0 \\ \mathcal{U}(0,1) & if\ L_{t+1} = 1 \end{cases}$$

Equation 7.2

Where δ is the delta function (1 at 0, 0 everywhere else,) and $\mathcal{U}$ is the uniform distribution. Intuitively, this means that if a jump has not occurred, the distribution stays exactly the same as in the last trial. If a jump has occurred, the distribution is uniform across all possible outcome probabilities.

The volatility is taken to be the log probability of witnessing a jump in the outcome probability at each trial: $P(\mathrm{L}_{t+1} = 1) = V$, where $V = exp(\mathrm{v})$, subject to $0 \leq V \leq 1$. v behaves similarly to a width parameter, such that a large value of v leads to a more uniform distribution over outcome probabilities, and a small value leads to a narrower distribution. The exponent V is taken so that integrals on v will be performed in log space, as is natural for a width parameter. We can then marginalise over the latent variable L_{t+1} to obtain:

$$P(g^A_{t+1}|g^A_t, V, S_{t+1} = A) = V \cdot \mathcal{U}(0,1) + (1 - V) \cdot \delta(g^A_{t+1} - g^A_t)$$

Equation 7.3

Note that unlike the original model of (Behrens et al., 2007) , due to reasons of computational cost I assume constant volatility rates (V) throughout the experiment. In the context of this experiment, this is a reasonable assumption as the volatility rate is indeed constant. However, the model can easily be extended to incorporate changing volatility.

The conditional distribution on g^B_{t+1}, while not used for an immediate prediction of the outcome when $S_{t+1} = A$, is tracked and calculated as

$$P(g^B_{t+1}|g^A_t, g^B_t, L_{t+1}, M_t, S_{t+1} = A) =$$

$$\begin{cases} \delta(g^B_{t+1} - ((1 - |M_t|) \cdot (g^B_t - 0.5) + M_t \cdot (g^A_t - 0.5) + 0.5)) & if\ L_{t+1} = 0 \\ \mathcal{U}(0,1) & if\ L_{t+1} = 1 \end{cases}$$

Equation 7.4

To understand the no-jump case of $L_{t+1} = 0$ intuitively, consider what happens when you plug in the different values of $M_t \in \{-1,0,1\}$. When $M_t = 0$ (i.e. the current A trial should not influence g^B_{t+1}, as A and B trials are independent), the term inside the delta function reduces to $g^B_{t+1} - g^B_t$, i.e. g^B does not change. When $M_t = 1$, the term inside the delta function reduces to $g^B_{t+1} - g^A_t$, i.e. g^B_{t+1} becomes equal to the previous distribution on A. When $M_t = -1$, the term inside the delta function reduces to $g^B_{t+1} - (1 - g^A_t)$, i.e. g^B_{t+1} is again fully determined by the previous distribution on A, but it will be exactly opposite (anticorrelated) to it.

Marginalising over the latent variable L_{t+1}, we obtain:

$$P(g^B_{t+1} | g^A_t, g^B_t, V, M_t, S_{t+1} = A) =$$

$$V \cdot \mathcal{U}(0,1) + (1 - V) \cdot \delta(g^B_{t+1} - ((1 - |M_t|) \cdot (g^B_t - 0.5) + M_t \cdot (g^A_t - 0.5) + 0.5))$$

Equation 7.5

If the correlation structure changes with time, optimal performance relies on the ability to profit from local correlation periods. I track the changing correlation, M_t, by again assuming the progression is Markovian, in a similar fashion to tracking the outcome probability:

$$P(M_{t+1} | M_t, K) = K \cdot \mathcal{U}(0,1) + (1 - K) \cdot \delta(M_{t+1} - M_t)$$

Equation 7.6

where $K = exp(k)$, subject to $0 \leq K \leq 1$, in a similar manner to V before. Large values of k will lead to a more uniform distribution over possible correlation values M_t, and would describe an environment where the correlation structure might change quickly, while small values of k will lead to narrower distributions, describing environments where the correlation structure could only change slowly or not at all. Hence, the parameter k can be thought of as the volatility of the correlation structure, or the amount of distrust in the current structure.

7.2.2.2 Inverse model - Inferring the outcome probabilities from past observations

The Bayesian learner's task is to use previous data (outcomes) to establish a belief about the statistics of the environment, in order to make an optimal decision in the next trial. This belief can be formulated as the *posterior* distribution, $P(unknowns|data)$, where in our case

the unknowns are the outcome probabilities, g^A and g^B, the volatility V, the correlation M and the volatility of the correlation K, and the data are past outcomes. The inference of the posterior distribution can be performed with Bayes rule:

$$P(unknowns|data) \propto P(data|unknowns) \cdot P(unknowns)$$

where $P(data|unknowns)$ is the *likelihood*, outlined in section 7.2.2.1 and **Figure 7.2**, and $P(unknowns)$ is the *prior* distribution over the unknown parameters. We can use Bayes rule to obtain the *joint* probability distribution of all the parameters at trial t, representing a complete picture of the statistics of the environment throughout the experiment:

$$P\left(g^A_{1,\dots t}, g^B_{1,\dots t}, V, M_{1,\dots t}, K \middle| y_{1,\dots t}, S_{1,\dots t}\right) \propto P\left(y_{1,\dots t} \middle| g^A_{1,\dots t}, g^B_{1,\dots t}, V, M_{1,\dots t}, K, S_{1,\dots t}\right) \cdot P\left(g^A_{1,\dots t}, g^B_{1,\dots t}, V, M_{1,\dots t}, K\right)$$

Equation 7.7

From the generative model in **Figure 7.2**, y_j and all other parameters are conditionally independent given g^A_j, g^B_j and S_j. Hence

$$P\left(y_{1,\dots t} \middle| g^A_{1,\dots t}, g^B_{1,\dots t}, V, M_{1,\dots t}, K, S_{1,\dots t}\right) = \prod_{j=1}^{t} P\left(y_j \middle| g^A_j, g^B_j, S_j\right)$$

Equation 7.8

By the conditional independencies dictated by the Markovian assumptions in the generative model, we obtain

$$\begin{aligned} P\left(g^A_{1,\dots t}, g^B_{1,\dots t}, V, M_{1,\dots t}, K\right) &= P(V)P(K)P(g^A_1)P(g^B_1)P(M_1) \\ &\cdot \prod_{j=2}^{t} P\left(g^A_j \middle| g^A_{j-1}, g^B_{j-1}, V, M_j, S_{j-1}\right) P\left(g^B_j \middle| g^B_{j-1}, g^A_{j-1}, V, M_j, S_{j-1}\right) \\ &\cdot P\left(M_j \middle| M_{j-1}, K\right) \end{aligned}$$

Equation 7.9

To obtain the posterior probability distribution for the unknown parameters at trial t, we substitute equations Equation 7.8 and Equation 7.9 into Equation 7.7 and marginalize over the history of g^A, g^B and M:

$$P\left(g_t^A, g_t^B, V, M_t, K \middle| y_{1,\dots t}, S_{1,\dots t}\right)$$
$$= \prod_{j=1}^{t} P\left(y_j \middle| g_j^A, g_j^B, S_j\right) P(V)P(K) \cdot \int \dots \int P(g_1^A)\, P(g_1^B)P(M_1)$$
$$\cdot \prod_{j=2}^{t} P\left(g_j^A \middle| g_{j-1}^A, g_{j-1}^B, V, M_j, S_{j-1}\right)$$
$$\cdot P\left(g_j^B \middle| g_{j-1}^B, g_{j-1}^A, V, M_j, S_{j-1}\right) P\left(M_j \middle| M_{j-1}, K\right)\, dg_{1,\dots t-1}^A dg_{1,\dots t-1}^B dM_{1,\dots t-1}$$

Equation 7.10

Similarly, we can obtain the posterior probability distribution for the unknown parameters at trial $t+1$:

$$P\left(g_{t+1}^A, g_{t+1}^B, V, M_{t+1}, K \middle| y_{1,\dots t+1}, S_{1,\dots t+1}\right)$$
$$= \prod_{j=1}^{t+1} P\left(y_j \middle| g_j^A, g_j^B, S_j\right) \cdot P(V)P(K)$$
$$\cdot \int \dots \int P(g_1^A)\, P(g_1^B)P(M_1) \prod_{j=2}^{t+1} P\left(g_j^A \middle| g_{j-1}^A, g_{j-1}^B, V, M_j, S_{j-1}\right)$$
$$\cdot P\left(g_j^B \middle| g_{j-1}^B, g_{j-1}^A, V, M_j, S_{j-1}\right) P\left(M_j \middle| M_{j-1}, K\right)\, dg_{1,\dots t}^A dg_{1,\dots t}^B dM_{1,\dots t}$$

Equation 7.11

This expression can be dramatically simplified by substituting in the posterior distribution of the parameters at the previous trial t from Equation 7.10:

$$P\left(g_{t+1}^A, g_{t+1}^B, V, M_{t+1}, K \middle| y_{1,\dots t+1}, S_{1,\dots t+1}\right)$$
$$= P(y_{t+1} | g_{t+1}^A, g_{t+1}^B, S_{t+1})$$
$$\cdot \iint \left(\int P(g_t^A, g_t^B, V, M_t, K) P(M_{t+1} | M_t, K)\ dM_t \right)$$
$$\cdot P(g_{t+1}^A | g_t^A, g_t^B, V, M_{t+1}, S_t) P(g_{t+1}^B | g_t^B, g_t^A, V, M_{t+1}, S_t) dg_t^A dg_t^B$$

Equation 7.12

Through this equation, the entire history of previous outcomes is summarised in the joint probability distribution - the current belief about the parameters g^A,g^B,V,M and K. This enables updating of that belief on a trial by trial basis, as can be seen by the absence of any data points previous to y_{t+1} or any parameters previous to trial t from the equation. We

have already found convenient forms for all other terms on the right-hand side of the equation in section 7.2.2.1.

I numerically calculated the integrals in the equation on a 8-dimensional grid (representing $g^A_{t+1}, g^B_{t+1}, g^A_t, g^B_t, V, M_{t+1}, M_t, K$, with integration over g^A_t, g^B_t and M_t). This leaves a 5-dimensional grid, representing the joint distribution $P\left(g^A_{t+1}, g^B_{t+1}, V, M_{t+1}, K \middle| y_{1,\dots t+1}, S_{1,\dots t+1}\right)$ that is carried over to the next trial. I set the joint distribution to be uniform at the beginning of each session of the experiment.

To obtain the model estimation of specific parameters, I marginalise over the rest of the parameters in the 5-dimensional joint distribution and compute the expected value of the marginal distribution. For example, to compute the model's current estimation of the *green* outcome probability of context A, I first compute the marginal distribution

$$P(g^A_{t+1}) = \int \dots \int P\left(g^A_{t+1}, g^B_{t+1}, V, M_{t+1}, K \middle| y_{1,\dots t+1}, S_{1,\dots t+1}\right)\, dg^B_{t+1} dV\, dM_{t+1} dK$$

Equation 7.13

and then take the expected value:

$$\widehat{g^A_{t+1}} = \int g^A_{t+1} P(g^A_{t+1})\, dg^A_{t+1}$$

Equation 7.14

It is also possible to use the maximum of the marginal distribution as its estimation, instead of the expected value, though in the results presented below I do not show that.

I tested the model with several configurations of the parameter space. For g^A and g^B, the parameter spaces used were either the $[0,1]$ interval discretised at 0.1 ("continuous outcome space"), or just the two true options for the outcome probabilities: $g \in \{0.05, 0.95\}$; For M I first used a "naïve" Bayesian model which doesn't know about the correlation ($M = 0$). I then used models that either had the space which included an optional independence of the two stimuli, as well as the two true correlation structures ($M \in \{-1, 0, 1\}$) (3 states model); lastly, I used the true parameter space ($M \in \{-1, 1\}$). The combinations of the possible parameter spaces for g and M result in 6 different models I have tried to fit the data with; $V \in \left[\frac{1}{40}, \frac{1}{4}\right]$, discretised to 10 bins on a logarithmic scale (equivalent to discretising $log\left(\left[\frac{1}{40}, \frac{1}{4}\right]\right)$ at 0.25 long equal intervals and then taking the

exponent); $K \in \left[\frac{1}{150}, \frac{1}{20}\right]$ discretised into 9 bins on a logarithmic scale (equivalent to discretising $log\left(\left[\frac{1}{40}, \frac{1}{4}\right]\right)$ at 0.25 long equal intervals and then taking the exponent). The denominators in the volatility parameters V and K are closely related to the average number of trials between switches (of outcome probabilities (V) or structure (K)), as explained in (Behrens et al., 2007).

7.2.3 Selector model

The estimations of the outcome probabilities, $\widehat{\mathrm{g}_{t+1}^{A}}$ and $\widehat{\mathrm{g}_{t+1}^{\mathrm{B}}}$, are used as input to a sigmoid function

$$P(choice = green) = \frac{1}{1 + e^{-\beta(\widehat{g^s} - 0.5)}}$$

Equation 7.15

where I fit the inverse temperature parameter β to subject's choices. To test for the model's accuracy of the predictions of subject's trial-by-trial choices, I also use a selector model where the sigmoid is actually a step function (i.e. $\beta \to \infty$) such that outcome probabilities above or below 0.5 always lead to a *green* or *blue* choice, respectively.

7.3 Results

7.3.1 Bayesian learner

The optimal agent is the one that makes the most efficient use of the history of information to track the outcome probabilities. In the task here, this includes inferring the correlation between outcome probabilities and using them, based only on the observed trial-by-trial outcomes. The problem of tracking two changing probabilities which might be correlated can be formalised as a probabilistic graphical model (**Figure 7.2**, see section 7.2.2 for an algebraic description).

In the following, the probabilities that the outcome following the presentation of objects A and B will be *green*[1] is denoted by g^A and g^B, respectively. These probabilities vary between trials, and are controlled by an outcome volatility parameter v which reflects how stable or volatile the outcome probabilities are (Behrens et al., 2007). Importantly, the outcome probabilities are also controlled by a correlation parameter M, which reflects the correlation between g^A and g^B: $M = 1$ means $g^A = g^B$ (i.e. positive correlation), $M = -1$ means $g^A = 1 - g^B$ (i.e. negative correlation) and $M = 0$ means g^A and g^B are independent. In addition, I use a "structure volatility" parameter k, that enables M to vary through time. The graphical description of the generative model is shown in **Figure 7.2**.

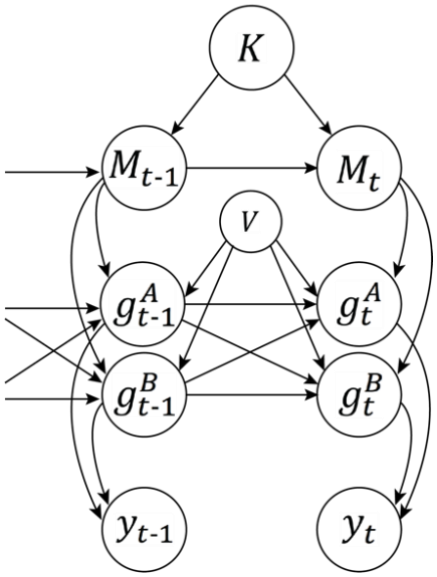


Figure 7.2 Graphical description of the correlated probability tracking problem.

Data, y is observed as a succession of trial outcomes, together with the stimulus in each trial $S \in \{A, B\}$. The Bayesian learner updates its estimates of the parameters g^A, g^B, v, M, k when it gets a new piece of information - an outcome of each trial, regardless of the type of stimulus. The parameter M determines the amount of update of g^B whenever an outcome

[1] This is analogous to learning whether an object will be rewarded or not in the task in chapters 5 and 6, with the choice between accepting and rejecting a trial (i.e. guessing if the trial is rewarded) being replaced by guessing whether the current outcome is green or blue. This change was made to avoid some of the confusions that arose with the previous paradigm, like the definitions of value and prediction errors.

on stimulus A is observed and vice versa. Crucially, the update equation relies only on the parameter estimates from the preceding trial and the latest trial outcome (of any stimulus) to determine decision and learning on the next trial (Equation 7.12). There is no need for the agent to store the entire history of recent outcomes.

I ran the Bayesian model with several different parameter spaces: outcome probabilities g^A,g^B had either a continuous $[0,1]$ space (discretised to 9 bins), or the true space of g^A,$g^B \in \{0.05,0.95\}$, while the structure variable M had 3 possible spaces: a structure-naïve model where $M = 0$; a 3- state space where $M \in \{-1,0,1\}$ (i.e. outcome probabilities can either be the opposite, uncorrelated or the same); and the true 2-state space, where $M \in \{-1,1\}$. The structure-naïve model served as a control, as well as a structure-naïve Rescorla-Wagner model (section 5.2.3).

Figure 7.3 (g^A,$g^B \in [0,1]$) and **Figure 7.4** (g^A,$g^B \in \{0.05,0.95\}$) show the performance of the models on the first 200 trials of an example schedule. The models with knowledge about the correlation structure successfully track it (**Figure 7.3**a, top) and use it to inform their estimations of the outcome probabilities (**Figure 7.3**a and **Figure 7.4**, middle and bottom, cyan lines). Unlike a structure-naïve Bayesian model (where M is set to 0, **Figure 7.3**a and **Figure 7.4**, magenta lines, as well as **Figure 7.3**b top row), following an outcome probability reversal, the models that know about the structure (**Figure 7.3**b bottom 4 rows) uses feedback on one stimulus to inform estimations about the other, enabling it a quicker response to the outcome probability reversal (**Figure 7.3**b, three leftmost columns). This only happens when the model has little uncertainty about the correlation structure (when the structure has been stable for a while). Following a reversal of the correlation structure, the model at first uses its strong belief about the old structure to (incorrectly) update estimations (**Figure 7.3**b, third and second columns from the right), until enough evidence is accumulated to "convince" the model that the structure has indeed changed (**Figure 7.3**b, rightmost columns).

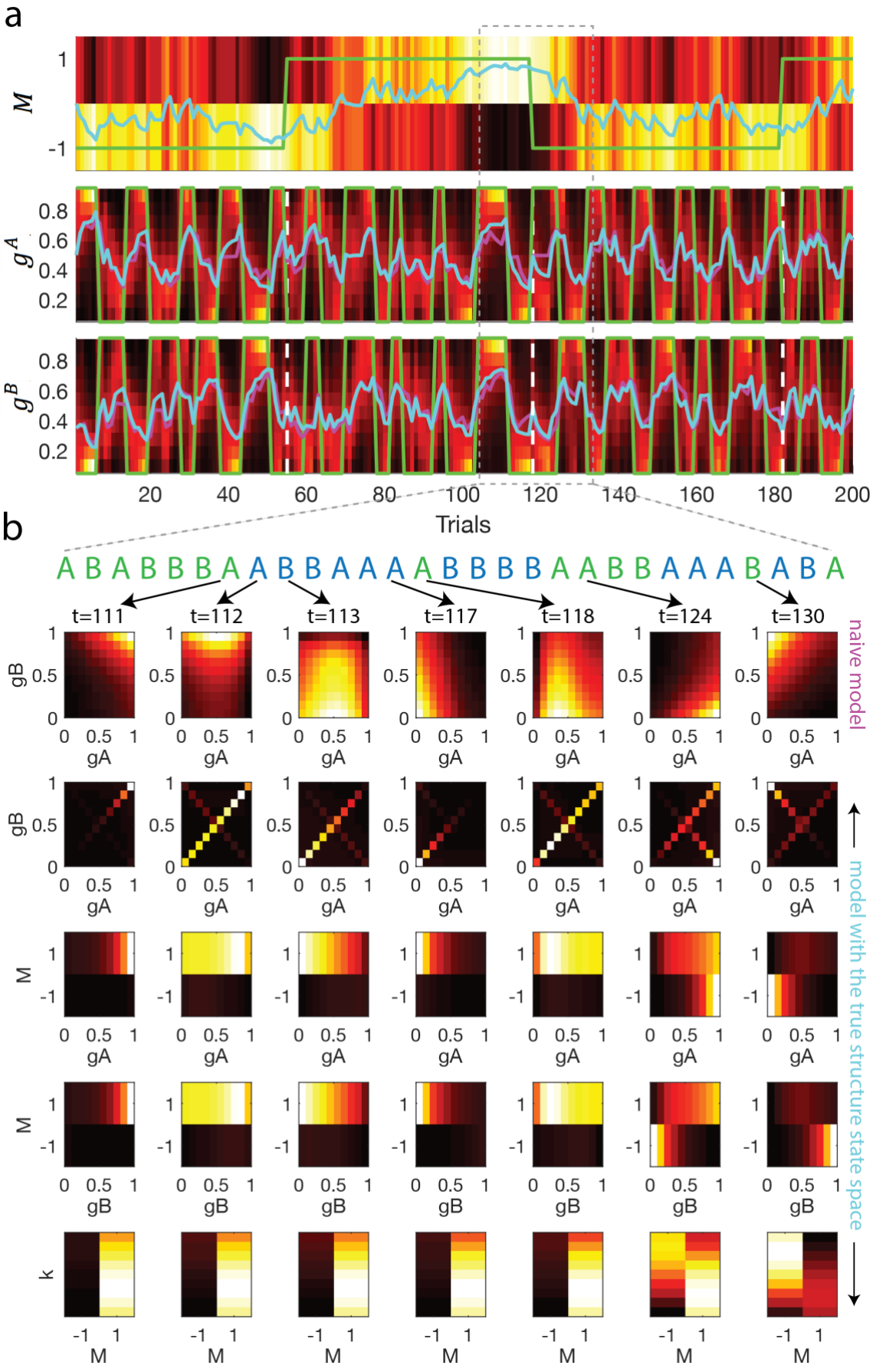

a
M
1
-1
g^A
0.8
0.6
0.4
0.2
g^B
0.8
0.6
0.4
0.2
20
40
60
80
100
120
140
160
180
200
Trials
b
A B A B B B A A B B A A A A B B B B A A B B A A A B A B A
t=111
t=112
t=113
t=117
t=118
t=124
t=130
naive model
model with the true structure state space
gB
gA
M
k
0
0.5
1
-1

Figure 7.3 Bayesian models with a (discretised) continuous outcome probability space (g^A,$g^B \in [0,1]$). a. Model estimates of the structure variable M (top) and the outcome probabilities of stimuli A (middle) and B (bottom) in the first 200 trials of an example schedule. Heatmaps: probability distributions (from the model with the true 2-state space $M \in \{-1,1\}$). Green lines: the true value of the variables. Cyan: expected value of the variables from g^A,$g^B \in [0,1]$, $M \in \{-1,1\}$ model. Magenta: expected value of the variables from the structure-naïve (g^A,$g^B \in [0,1]$, $M = 0$). White dashed lines: structure reversals. Note that the estimate from the model that knows about the structure (cyan) tracks the true probability quicker after outcome reversals but is slower to adapt after a structure reversal. (see **Figure 7.4** for a clearer demonstration of this point) **b.** Schedule (stimuli A/B coloured in the colour of the outcome (green/blue) following them) and posterior marginal distributions for trials 105-133 (marked with grey dashed square in **a.**). Top row: outcome probabilities from the structure-naïve model (g^A,$g^B \in [0,1]$, $M = 0$). Rows 2-5: Marginal posterior distributions of different pairs of variables from g^A,$g^B \in [0,1]$, $M \in \{-1,1\}$ model.

7.3.2 Subjects' behaviour

I piloted the current task paradigm on 5 subjects, recruited online. I also ran the 7 models described above (a Rescorla-Wagner model and 6 Bayesian models with different combinations of parameter spaces for the outcome probabilities (g^A,g^B) and the structure variable (M)).

To test how well the models predicted a subject's behaviour, I calculated the negative log likelihood of each model, given the subject's choices in the last half of the trials, after the subject had a chance to experience the task structure for (and several structure reversals) for the first half (**Table 7.1**). I also tested whether the models predicted the subject's choices by directly comparing them with the choices made by the model (where if the expected value of the outcome probability g was above 0.5, the model's choice was taken to be *green*, and otherwise it was taken to be *blue*, Table 7.2). As in most trials most models make the same predictions, I also calculated this measure on a subset of "difficult" trials, where the predictions of all models were not congruent (Table 7.2, values in brackets). These were usually trials after experiencing a reversal. The best fitting model for subjects 3,4 and 5 was the Bayesian model with the true parameter space, suggesting these subjects have learnt the task structure and used it to guide their choices. The best fitting models for subjects 1 and 2 were structure-naïve models (either Rescorla-Wagner or Bayesian), suggesting these subjects did not use the correlation structure to guide their choices.

model / subj.	RW	$g \in \{.05,.95\}$ $M = 0$	$g \in [0,1]$ $M = 0$	$g \in \{.05,.95\}$ $M \in \{-1,1\}$ (true space)	$g \in [0,1]$ $M \in \{-1,1\}$	$g \in \{.05,.95\}$ $M \in \{-1,0,1\}$	$g \in [0,1]$ $M \in \{-1,0,1\}$
1	138.4	138.5	138.5	138.6	138.6	138.6	138.6
2	73.6	79.4	82.3	84.8	78.1	81.4	77.3
3	109.9	103.8	110.5	92.8	97.2	93.4	99.8
4	99.6	93.8	108.5	60.0	78.7	65.4	84.5
5	130.0	124.1	129.4	120.2	125.7	120.0	126.1

Table 7.1. Negative log likelihoods in the last 200 trials for 5 subjects. Lower numbers indicate better fit of the model. Each row is a different subject. Cells are coloured with the same colours as the estimates of variables in Figure 7.3a ($g \in [0,1]$, columns 3 and 5) and Figure 7.4 ($g \in \{.05,.95\}$, columns 2 and 4).

model / subj.	RW	$g \in \{.05,.95\}$ $M = 0$	$g \in [0,1]$ $M = 0$	$g \in \{.05,.95\}$ $M \in \{-1,1\}$ (true space)	$g \in [0,1]$ $M \in \{-1,1\}$	$g \in \{.05,.95\}$ $M \in \{-1,0,1\}$	$g \in [0,1]$ $M \in \{-1,0,1\}$
1	54.5 (58.4)	50.50 (50.5)	50.5 (50.5)	49 (47.5)	53.5 (56.4)	48.5 (46.5)	54.5(58.4)
2	87.5 (75.8)	87.5 (75.8)	87.5 (75.7)	79 (24.2)	84 (54.5)	80.5 (33.3)	85.5(63.6)
3	75.5 (35.1)	76 (37.8)	76 (37.8)	81 (64.9)	75.5 (35.1)	79.5 (56.8)	76.5 (40.5)

4	79.5 (24.3)	80 (27.0)	80 (27.0)	89.5 (78.4)	80 (27.0)	87.5 (67.6)	79 (21.6)
5	64 (34.3)	65 (40.0)	65 (40.0)	69.5 (65.7)	65.5 (42.9)	70.5 (71.4)	75.5 (42.9)

Table 7.2. Predictions of model in the last 200 trials. Given as % of model choices which are the same as the subject's choices. Numbers in brackets are the % of correct predictions in a subset of "difficult" trials – trials for which there was some disagreement between choices made by any of the 7 models. Each row is a different subject. Cells are coloured with the same colours as the estimates of variables in Figure 7.3a ($\boldsymbol{g} \in [\mathbf{0}, \mathbf{1}]$, columns 3 and 5) and Figure 7.4 ($\boldsymbol{g} \in \{.\mathbf{05}, .\mathbf{95}\}$, columns 2 and 4).

7.3.3 Supplementary figures

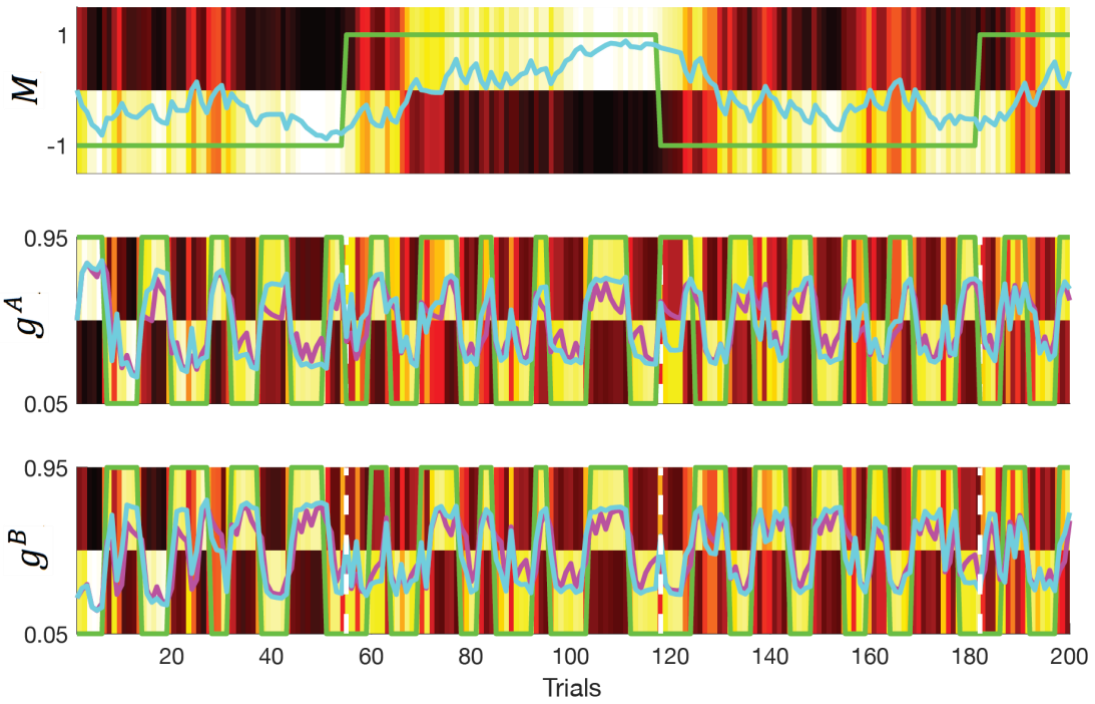


Figure 7.4 Parameter estimates for models with the true, 2-state parameter space for the outcome probabilities. Same figure like Figure 7.3a but for the models with g^A,$g^B \in \{0.05,0.95\}$. Model estimates of the structure variable M (top) and the outcome probabilities of stimulus A (middle) and B (bottom) in the first 200 trials of an example schedule. Heatmaps: probability distributions (from the model with the true 2-state outcome probability space g^A,$g^B \in \{0.05,0.95\}$ and the true 2-state space $M \in \{-1,1\}$). Green lines: the true value of the variables. Cyan: expected value of the variables from g^A,$g^B \in \{0.05,0.95\}$, $M \in \{-1,1\}$ model. Magenta: expected value of the variables from the structure-naïve (g^A,$g^B \in \{0.05,0.95\}$, $M = 0$). White dashed lines: structure reversals. Note that the estimate from model that knows about the structure (cyan) tracks the true probability quicker after outcome reversals but is slower to adapt after a structure reversal.

7.4 **Discussion**

In this chapter I presented an RL task where the relational structure dynamically changes. I described a Bayesian learner that infers the current structure only from outcome observations, enabling it to use the outcome information in an optimal manner. Preliminary data show that the behaviour of some subjects can indeed be explained by this model, suggesting these subjects perform the structural inference in a close-to-optimal way and use it to inform their choices.

As the reader can imagine, these are only preliminary results in a more ambitious project. Although the behavioural capability of some subjects to perform the structure inference and use it to guide behaviour is non-trivial, the question I am really interested in is the nature of structural representations in a dynamic environment and the structural inference process. For example, it is reasonable to think that the EC structural representation described in chapter 5 will also be used when the structure changes dynamically. The Bayesian model is a powerful tool for the investigation of such representations with model-based fMRI. For example, an easily testable prediction is that the strength of the representation will vary as a function of the uncertainty over the M structure parameters. Another possible future direction is to study the structural volatility signal: in the current task, both the average outcome probability volatility and the structural volatility were constant. (Behrens et al., 2007) showed that the outcome probability volatility was tracked by the anterior cingulate cortex. It will be interesting to test whether a dissociation can be found between that signal and the (hierarchically higher-level) structural volatility.

The results from the structure remapping experiment in physical space by (Boccara et al., 2019), together with my results from chapter 5 generate strong predictions regarding the mechanism allowing structural inference in the task presented here: that the two structures are represented in different entorhinal maps, with the inference process over these maps having a signature in the form of the flipping between the representations of the two maps (**Figure 3.2**c,d). While this temporally fast signature is not accessible to investigation with fMRI (though see (Schuck and Niv, 2019) for a possible way forward with fMRI), it might be possible to index it with MEG, using techniques similar to the ones used by (Kurth-Nelson et al., 2016; Liu et al., 2019). I hope to test these predictions over the coming year.

8 GENERAL DISCUSSION AND FUTURE DIRECTIONS

In this thesis, I aimed to contribute to a growing body of literature bringing together the research fields of cognitive maps and reinforcement learning - fields that largely progressed in parallel until the last decade. Through interpreting physical space within the RL (MDP) framework in combination with recent suggestions regarding the computational features of grid cells, I presented a novel planning algorithm that is both flexible and efficient, and that can be extended to arbitrary, non-spatial environments – intuitive planning. Taking inspiration from remapping experiments in physical space, I investigated neural representations of the structure of RL tasks. I found that the entorhinal cortex that codes for the structure of physical space, also coded for the structure of an abstract RL task. In addition, I used the remapping paradigm to suggest a novel interpretation of a heavily studied RL signal - the reward prediction error, which I also found to be sensitive to the task structure. These findings, together with several other results that fit well with the existing literature, have strong implications regarding the brain's mechanisms for generalisation – a fundamental feature of intelligence.

8.1 A unifying computational framework: hierarchical embedding spaces for task states

In this section I cast the analyses presented in chapters 5 and 6 as taking averages over different levels of a hierarchical system for representing and navigating the task state space. First, it is important to clarify how the "correct" underlying state space of the task looks (though it is impossible to know from the data whether this is actually the state-space the subjects are using, and it is possible to solve the task with different state spaces). In each block, there are two possible states for the two related stimuli: for example in -Corr blocks, trials where the probability of a "good" outcome is high for cue A and low for cue B belong to one state (A high B low - AhBl) while the rest of the A or B trials belong to the other state (AlBh). Analogously, in +Corr blocks, the two possible states in A or B trials are AhBh and AlBl. With the same notation, the control cue (C) trials have two possible states: Ch or Cl,

regardless of the relational structure, resulting in 4 possible states in each block (-Corr blocks: AhBl, AlBh, Ch, Cl; +Corr blocks: AhBh, AlBl, Ch, Cl).

Previous work has suggested that the OFC codes for the current position within the task space (Schuck et al., 2016; Wilson et al., 2014). Here, that would mean indicating which of the 4 states in each block is currently relevant. This theory predicts that OFC representations should be similar for trials of the same state but different between different states. Unfortunately, we could not directly test this prediction due to low number of trials of each state in each block. However, our within-block credit assignment effect (section 6.3.1.2.1) could be viewed as a comparison between the average representation of the two AB states and the average representation of the two C states. As half of the averaged trials (those of the same state) should have a similar representation within AB trials and the same within C trials, but not across these groups of trials, the predicted OFC result is a higher representational similarity between A to B trials than A to C and B to C trials – as indeed we observed.

This "current position in current task map" signal must be informed by representations carrying information about the correct current relational structure, permitting construction of correct state spaces and latent states which enable generalisation. Throughout this thesis I have highlighted, through both theoretical and experimental arguments, why in physical space grid cells are a natural candidate for such a representation. However, the grid cell population has also been shown to code for the current location of an animal (Stemmler et al., 2015) and to enable vector navigation from this location to a target location (Bush et al., 2015). Following the hierarchical computational framework presented in (Whittington et al., 2019), I suggest that here EC (and perhaps vmPFC) might also code for the latent position in the task space, but it does so in the same neuronal embedding subspace (manifold) for the same relational structures, and different neuronal embedding subspaces for different relational structures. This means that representations of task states of the same relational structure (AhBl and AlBh in -Corr blocks, AhBh and AlBl in +Corr blocks, regardless of what the stimuli A and B are) lie on the same EC "structure manifold" and hence *can* (though might not) have similar representations. Representations of task states of different relational structures lie on different EC neuronal manifolds, and hence necessarily have dissimilar representations. The EC structure representation reported in this thesis (section 5.3.2) can then be viewed as arising from a similarity between latent position representations, which are "recycled" for states with different sensory instantiations (stimuli sets) of the same relational

structure. This similarity might be strong enough to mitigate the limitation of the analysis where each condition in the RDM is an average of the trials from the two states a stimulus in a particular block can be tied to. In addition, I note that the (weak) effects we find in EC in the within-block structural credit assignment (**Figure 6.2**) are consistent with a representation of the current state.

The exact same logic applies to the prediction error x structure effect in vmPFC, PCC and vlPFC (section 6.3.2): these results are expected if learning signals on task states are embedded in the same manifold for one structure but in a different manifold for the other structure. Such representations might then inform the construction of the correct state space in OFC, as is hinted by the correlation between the effects (**Figure 6.5**).

Importantly, I stress that the notion of a "structure manifold" does not conflict with an explicit structural representation but is rather a nuanced version of it in which the current location is coded as well as the structure. This interpretation can explain how the same EC cells can be active in more than one structure (e.g. (Boccara et al., 2019)), as the activity of a single cell can be very similar even if the manifold spanned by the population activity is different.

8.2 **Open questions and future work**

There are many open questions that arise from the work in this thesis. Here I will review the ones I find most interesting.

Grid cells in abstract tasks. The structural representation results presented in chapter 5 point to the EC as an area that codes for structural knowledge in an RL task. It is also known to harbour grid cells, which code for the structure of 2D spaces. However, it is not known what the activity of grid cells looks like in such tasks like the one used here. If grid cells maintain their computational properties in such tasks, we can hope for a detailed and mechanistic understanding of these abstract structural representations – for example in the form of the eigenvectors of the task transition structure (section 4.2.3). However, it is no doubt possible that different entorhinal populations code for different structures in different task. Still, the results presented in this thesis predict that whatever representation might be found in the EC cells in these abstract tasks, it will be a low-dimensional projection of the

structure of the task. These questions will hopefully be addressable through recordings in animals solving both spatial and abstract tasks.

Origins of the PE x structure effect. Similarly, it is not quite clear what the mechanism is from which the PE x structure interaction effect arises. Do the mPFC spatial grid cells (Jacobs et al., 2013) also code for the structure in our task? We did not find evidence for direct abstract coding of the task in mPFC as we did in EC. Perhaps mPFC grid cells (or the cells that code for the structure in my task, if they are not the same as spatial grid cells) modulate reward-sensitive vmPFC cells, giving rise to differential PE coding across structures? If so, what are the mechanistic properties of such modulation – e.g. are there separate PE-coding populations for different structures?

The relations between state-spaces, states and behaviour. Another important question, which I tried to speculate about in the previous section, is the relationship between abstract structure representations and current state representations, and the relationship between these and behaviour. (Schuck and Niv, 2019) showed that while the current state of an abstract task can be decoded in both HPC and OFC, only the OFC representation is important for behaviour. Indeed, I also observed a trend in hippocampus (though far from significant) for the latent variable analysis (section 6.3.1.2.1, data not shown), which can be viewed as a proxy to current state representation (section 8.1). The (medial) EC structural representation might be, as it is in physical space, the structural input to a hippocampal representation of the conjunction of stimulus and structure (Manns and Eichenbaum, 2006; Whittington et al., 2019). This hippocampal representation might in turn be important for constructing/solidifying the online, behaviourally relevant OFC current state representation (Schuck and Niv, 2019; Wikenheiser et al., 2017) – a shadow of which I see in my latent variable analysis (section 3.3.1.2.1). Unfortunately, the behavioural measures in the task presented in chapter 5 were not sensitive enough to find correlations with neural effects (as almost all subjects were at ceiling). Future work, with more sensitive behavioural measures, might be able to answer some of these questions.

Other space-like coding features in RL tasks. Physical space has proved a useful inspiration in this work, as well as many other studies, due to our good understanding of the neural representations underlying it. It is possible that some other representational feature of physical space will also transfer to abstract domains. For example, in both the EC (Stensola et al., 2012) and HPC (Strange et al., 2014), functional gradients of firing field sizes can be found. Hierarchies are also prevalent and very useful in RL tasks (Botvinick et al., 2009). It

is possible that the same functional gradients from space will also apply in hierarchical RL tasks.

Structure inference: In chapter 7, I presented an experiment where the (simple) relational structure of an RL task changed dynamically and needed to be inferred on-the-fly by the subject. The neural signals underlying the online inference of structure have not been explored. For example, what is the relationship between the reward volatility signals in the ACC (Behrens et al., 2007) and the higher-level "structure volatility"? What are the mechanisms underlying the inference process, and are they similar to the flickering mechanisms discovered in rats performing structural inference (Boccara et al., 2019)? The model and task presented in chapter 7 will hopefully prove useful in addressing these questions.

Structure learning: The results presented in this thesis provide a suggestion to a key part of the reinforcement learning problem – how the state space is represented. However, it does not address the question of the initial learning of this representation. Indeed, how, and when, do we learn a novel structure? Is it in fact possible to learn new structures?

Some recent work suggests a mechanism by which canonical structural forms that are abundant in nature (Kemp and Tenenbaum, 2008) serve as a basis sets for structural representations, such that in novel tasks the structural form is inferred and composed from the known "recycled" structural representations (Mark et al., 2019). Taken to the extreme, this viewpoint suggests that there are a limited number of "natural" structures, probably learnt during childhood (for example the structures "positive correlation" and "negative correlation", as well as "2D spaces", "1D lines" (e.g. time), cluster community structures etc.), which then constrain adult experiences through their (fitted) combinations. The question of structure learning, and the extent to which it can happen, is indeed a remarkably open one.

APPENDIX – NEURORISK

I include here what was probably the best idea I had during my DPhil. In the run-up to the lab retreat of 2018 Tim Muller and I developed an educational game called NeuroRisk, where players compete for brain domination by learning about neuroanatomy and discovering cool facts about the brain. Neurorisk is inspired by the board game Risk, a strategy board game where players compete for world domination through the conquest of territories on the world map, with battles for territories being decided through dice rolls. We adapted Risk to NeuroRisk, where instead of the world, the map is the human cortex. People are split into multiple teams (of ~5 people) that are each given a mission, such as "conquer the temporal lobe". Each team will start the game with their soldiers randomly distributed throughout the brain. Battles for brain regions are conducted by both defending and attacking teams presenting some pitch (e.g. interesting or funny fact) about that brain region in question, with the winning side decided by popular vote. The rules are below.

Importantly, attacks can only be made to brain regions that are functionally connected. To determine functional connectivity, we downloaded and used the parcellation by (Glasser et al., 2016) along with their HCP Workbench viewer. We clustered brain regions together to reduce the number of parcels from ~200 to ~50, resulting in our own map of human cortex, i.e. the board of the game. Using the HCP Workbench, we could then online – during the game – determine functional connectivity, i.e. between which brain regions (or territories) attacks could be made.

We hope to publish the game on bioRxiv, and we believe this will be by far our greatest and most impactful contribution to the neuroscience community from our DPhil period.

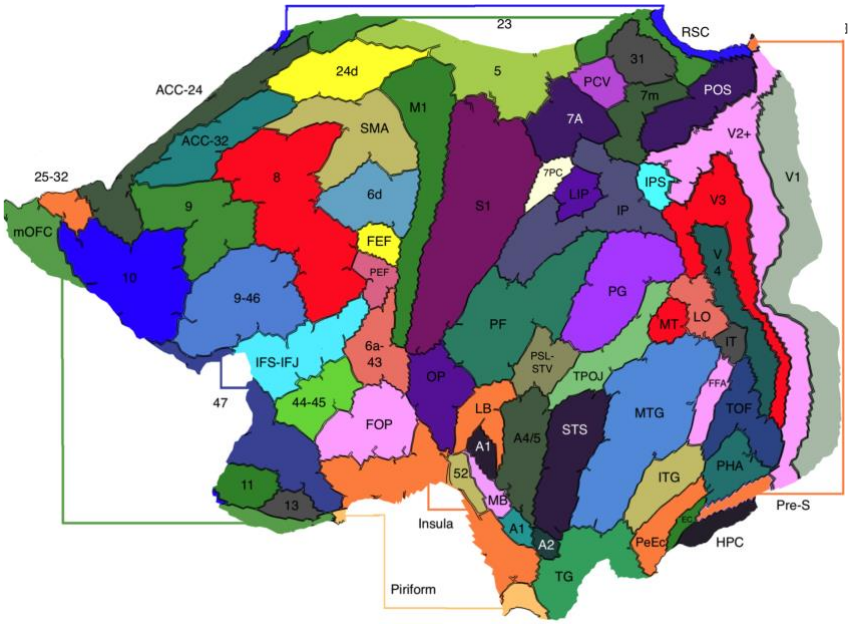


Figure 0.1 Map of the human cortex, i.e. the board in NeuroRisk.



Figure 0.2 Neurorisk in progress. Below: Board and soldiers (skittles) on it. **Above**: People searching for facts on their laptops. Age of Empires battle music is playing in the background.

NeuroRisk Rules:

The world:

- Human cortex.

Teams & soldiers:

- Five teams.
- Each team will start with a soldier on ~ 1/5 of the territories.
- Plus more soldiers to distribute as you please, according to your mission.
- At the end of each go, you can redistribute your soldiers.
- You can only eat skittles of your team colour (this is not recommended).

Your turn and attacking:

- At the start of your turn, you will receive more new soldiers; the number will depend on how many regions you conquered on the last go & number of current territories; According to: ceil(current_territories/3) + randi(conquered_territories)
- You can keep attacking regions for as long as you want when it's your turn.
- Challenge a brain region you wish to attack, and state from which region you are attacking (that you currently occupy).
- You must be functionally connected to the region you are attacking from your current region. Being adjacent in territories is NOT enough to attack.
- Attacker states how many soldiers they are attacking with. They must have at least the stated number + 1 on the region from which they are attacking.
- To attack, the attacker pitches something interesting or cool about that region.
- Defender responds with their own interesting or cool fact about that region.
- Everyone else that is not in either of those two teams, votes on which pitch wins.
- The loser loses the number of soldiers that the attack was made with.
- If the defender now has no soldiers left on that region, the attacker takes that region.
- The attacker cannot leave their original territory empty.
- At the end of your turn, you can redistribute your soldiers (but must keep at least one on each of your occupied regions).

- Optional rule – To get around the problem of regions no one knows, there can be dice attacks:
- Attacker can choose to attack with dice, but with bad odds.
- If attacker chooses to attack with facts, defender can choose dice, but they will have bad odds.
- Whether it is a facts or dice attack is decided before any facts are pitched.
- Therefore the game will not halt at these regions, but there is still incentive to pitch facts.

REFERENCES

Acuña, D.E., Schrater, P., 2010. Structure learning in human sequential decision-making. PLoS Comput. Biol. 6, e1001003.

Adams, C.D., 1982. Variations in the Sensitivity of Instrumental Responding to Reinforcer Devaluation. Q. J. Exp. Psychol. Sect. B 34, 77–98.

Adams, C.D., Dickinson, A., 1981. Instrumental responding following reinforcer devaluation. Q. J. Exp. Psychol. Sect. B 33, 109–121.

Akam, T., Costa, R., Dayan, P., 2015. Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. PLoS Comput. Biol. 11, 1004648.

Ambrose, R.E., Pfeiffer, B.E., Foster, D.J., 2016. Reverse Replay of Hippocampal Place Cells Is Uniquely Modulated by Changing Reward. Neuron 91, 1124–1136.

Aronov, D., Nevers, R., Tank, D.W., 2017a. Mapping of a non-spatial dimension by the hippocampal-entorhinal circuit. Nature 543, 719–722.

Aronov, D., Nevers, R., Tank, D.W., 2017b. Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. Nature 543, 719–722.

Atallah, H.E., Lopez-Paniagua, D., Rudy, J.W., O'Reilly, R.C., 2007. Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. Nat. Neurosci. 10, 126–131.

Averbeck, B.B., Costa, V.D., 2017. Motivational neural circuits underlying reinforcement learning. Nat. Neurosci. 20, 505–512.

Baldassano, C., Hasson, U., Norman, K.A., 2018. Representation of Real-World Event Schemas during Narrative Perception. J. Neurosci. 38, 9689–9699.

Balleine, B.W., 2005. Neural bases of food-seeking: Affect, arousal and reward in corticostriatolimbic circuits. Physiol. Behav. 86, 717–730.

Balleine, B.W., Dickinson, A., 1991. Instrumental Performance following Reinforcer Devaluation Depends upon Incentive Learning. Q. J. Exp. Psychol. Sect. B 43, 279–296.

Balleine, B.W., Dickinson, A., 1998. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology 37, 407–419.

Balleine, B.W., O'Doherty, J.P., 2010. Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. Neuropsychopharmacology 35, 48–69.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M.J., Degris, T., Modayil, J., Wayne, G., Soyer, H., Viola, F., Zhang, B., Goroshin, R., Rabinowitz, N., Pascanu, R., Beattie, C., Petersen, S., Sadik, A., Gaffney, S., King, H., Kavukcuoglu, K., Hassabis, D., Hadsell, R., Kumaran, D., 2018. Vector-based navigation using grid-like representations in artificial agents. Nature 557, 429–433.

Bao, X., Gjorgieva, E., Shanahan, L.K., Howard, J.D., Kahnt, T., Gottfried, J.A., 2019. Grid-like Neural Representations Support Olfactory Navigation of a Two-Dimensional Odor Space. Neuron 102, 1066-1075.e5.

Baram, A.B., Muller, T.H., Behrens, T.E.J., 2017. Intuitive planning: global navigation of cognitive maps with grid-like global representations. In: Proceedings of the Computational and Systems Neurosciscience Meeting.

Baram, A.B., Muller, T.H., Whittington, J.C.R., Behrens, T.E.J., 2018. Intuitive planning: global navigation through cognitive maps based on grid-like codes. bioRxiv 421461.

Barron, H.C., Dolan, R.J., Behrens, T.E.J., 2013. Online evaluation of novel choices by simultaneous representation of multiple memories. Nat. Neurosci. 16, 1492–1498.

Barron, H.C., Garvert, M.M., Behrens, T.E.J., 2016. Repetition suppression: A means to index neural representations using BOLD? Philos. Trans. R. Soc. B Biol. Sci.

Barry, C., Ginzberg, L.L., O'Keefe, J., Burgess, N., 2012. Grid cell firing patterns signal environmental novelty by expansion. Proc. Natl. Acad. Sci. 109, 17687–17692.

Bartra, O., McGuire, J.T., Kable, J.W., 2013. The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage 76, 412–427.

Beckmann, M., Johansen-Berg, H., Rushworth, M.F.S., 2009. Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. J. Neurosci. 29, 1175–90.

Behrens, T.E.J., Muller, T.H., Whittington, J.C.R., Mark, S., Baram, A., Stacenhfeld, K.L., Kurth-Nelson, Z., 2018a. What is a cognitive map ? Organising knowledge for flexible behaviour. BioRxiv.

Behrens, T.E.J., Muller, T.H., Whittington, J.C.R., Mark, S., Baram, A.B., Stachenfeld, K.L., Kurth-Nelson, Z., 2018b. What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. Neuron.

Behrens, T.E.J., Woolrich, M.W., Walton, M.E., Rushworth, M.F.S., 2007. Learning the value of information in an uncertain world. Nat. Neurosci. 10, 1214–21.

Beierholm, U.R., Anen, C., Quartz, S., Bossaerts, P., 2011. Separate encoding of model-based and model-free valuations in the human brain. Neuroimage 58, 955–962.

Bellman, R., 1957. Dynamic Programming. Princeton University Press.

Blodgett, H.C., 1929. The effect of the introduction of reward upon the maze performance of rats. Univ. Calif. Publ. Psychol. 113–134.

Boccara, C.N., Nardin, M., Stella, F., O'Neill, J., Csicsvari, J., O'Neill, J., Csicsvari, J., 2019. The entorhinal cognitive map is attracted to goals. Science (80-.). 363, 1443–1447.

Bongioanni, A., Folloni, D., Verhagen, L., Sallet, J., Klein-Flügge, M.C., Rushworth, M.F.S., 2020. Novel choices in macaques: activation and disruption of a neural mechanism for value integration. under Rev.

Bonnevie, T., Dunn, B., Fyhn, M., Hafting, T., Derdikman, D., Kubie, J.L., Roudi, Y., Moser, E.I., Moser, M.-B., 2013. Grid cells require excitatory drive from the hippocampus. Nat. Neurosci. 16, 309–317.

Boorman, E.D., Behrens, T.E.J., Woolrich, M.W., Rushworth, M.F.S., 2009. How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. Neuron 62, 733–743.

Boorman, E.D., Rajendran, V.G., O'Reilly, J.X., Behrens, T.E.J., 2016. Two Anatomically and Computationally Distinct Learning Signals Predict Changes to Stimulus-Outcome Associations in Hippocampus. Neuron 89, 1343–1354.

Boorman, E.D., Rushworth, M.F.S., Behrens, T.E., 2013. Ventromedial prefrontal and anterior cingulate cortex adopt choice and default reference frames during sequential multi-alternative choice. J. Neurosci. 33, 2242–53.

Bostock, E., Muller, R.U., Kubie, J.L., 1991. Experience-dependent modifications of hippocampal place cell firing. Hippocampus 1, 193–205.

Botvinick, M.M., 2012. Hierarchical reinforcement learning and decision making. Curr. Opin. Neurobiol. 22, 956–962.

Botvinick, M.M., Niv, Y., Barto, A.G., 2009. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. Cognition 113, 262–280.

Bouret, S., Richmond, B.J., 2010. Ventromedial and Orbital Prefrontal Neurons Differentially Encode Internally and Externally Driven Motivational Values in Monkeys. J. Neurosci. 30, 8591–8601.

Bowman, C.R., Zeithamova, D., 2018. Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. J. Neurosci. 38, 2811–17.

Bradfield, L.A., Dezfouli, A., van Holstein, M., Chieng, B., Balleine, B.W., 2015. Medial Orbitofrontal Cortex Mediates Outcome Retrieval in Partially Observable Task Situations. Neuron 88, 1268–1280.

Brogden, W.J., 1939. Sensory pre-conditioning. J. Exp. Psychol. 25, 323–332.

Bunsey, M., Elchenbaum, H., 1996. Conservation of hippocampal memory function in rats and humans. Nature 379, 255–257.

Bush, D., Barry, C., Manson, D., Burgess, N., 2015. Using Grid Cells for Navigation. Neuron 87, 507–520.

Butler, W.N., Hardcastle, K., Giocomo, L.M., 2019. Remembered reward locations restructure entorhinal spatial maps. Science (80-.). 363, 1447–1452.

Camille, N., Tsuchida, A., Fellows, L.K., 2011. Double dissociation of stimulus-value and action-value learning in humans with orbitofrontal or anterior cingulate cortex damage. J. Neurosci. 31, 15048–52.

Carelli, R.M., 2002. The nucleus accumbens and reward: neurophysiological investigations in behaving animals. Behav. Cogn. Neurosci. Rev. 1, 281–96.

Cavada, C., Compañy, T., Tejedor, J., Cruz-Rizzolo, R.J., Reinoso-Suárez, F., 2000. The anatomical connections of the macaque monkey orbitofrontal cortex. A review. Cereb. Cortex 10, 220–242.

Cavada, C., Schultz, W., 2000. The Mysterious Orbitofrontal Cortex. Foreword. Cereb. Cortex 10, 205–205.

Chase, H.W., Kumar, P., Eickhoff, S.B., Dombrovski, A.Y., 2015. Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. Cogn. Affect. Behav. Neurosci. 15, 435–459.

Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., Uchida, N., 2012. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature 482, 85–88.

Cohen, N.J., Eichenbaum, H., 1993. Memory, Amnesia, and the Hippocampal System. Cambridge: MIT Press.

Cohen, N.J., Squire, L.R., 1980. Preserved Learning and Retention of Pattern-Analyzing Skill in Amnesia : Dissociation of Knowing How and Knowing that Author (s): Neal J . Cohen and Larry R . Squire. Science (80-.). 210, 207–210.

Cole, R.P., Barnet, R.C., Miller, R.R., 1995. Temporal encoding in trace conditioning. Anim. Learn. Behav. 23, 144–153.

Collin, S.H.P., Milivojevic, B., Doeller, C.F., 2015. Memory hierarchies map onto the hippocampal long axis in humans. Nat. Neurosci.

Colwill, R.M., Rescorla, R.A., 1985. Postconditioning devaluation of a reinforcer affects instrumental responding. J. Exp. Psychol. Anim. Behav. Process. 11, 120–132.

Constantinescu, A.O., OReilly, J.X., Behrens, T.E.J., 2016. Organizing conceptual knowledge in humans with a gridlike code. Science (80-.). 352, 1464–1468.

Corbit, L.H., Balleine, B.W., 2003. The role of prelimbic cortex in instrumental conditioning. Behav. Brain Res. 146, 145–157.

Corbit, L.H., Balleine, B.W., 2011. The general and outcome-specific forms of Pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell. J. Neurosci. 31, 11786–94.

Corbit, L.H., Muir, J.L., Balleine, B.W., 2001. The role of the nucleus accumbens in instrumental conditioning: Evidence of a functional dissociation between accumbens core and shell. J. Neurosci. 21, 3251–60.

Corneil, D.S., Gerstner, W., 2015. Attractor Network Dynamics Enable Preplay and Rapid Path Planning in Maze–like Environments.

Cushman, F., Morris, A., 2015. Habitual control of goal selection in humans. Proc. Natl. Acad. Sci. 112, 13817–13822.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. Neuroimage 9, 179–194.

Damasio, H., Grabowski, T., Frank, R., Galaburda, A., Damasio, A., 1994. The return of Phineas Gage: clues about the brain from the skull of a famous patient. Science (80-.). 264, 1102–1105.

Danjo, T., Toyoizumi, T., Fujisawa, S., 2018. Spatial representations of self and other in the hippocampus. Science (80-.). 359, 213–218.

Daw, N.D., 2016. Reinforcement Learning. In: Arbib, M., Bonaiuto, J.J. (Eds.), From Neuron to Cognition via Computational Neuroscience. MIT Press.

Daw, N.D., Dayan, P., 2014. The algorithmic anatomy of model-based evaluation. Philos. Trans. R. Soc. B Biol. Sci. 369, 20130478–20130478.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron 69, 1204–1215.

Daw, N.D., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. 8, 1704–1711.

Daw, N.D., Tobler, P.N., 2014. Value Learning through Reinforcement: The Basics of Dopamine and Reinforcement Learning. Neuroeconomics 283–298.

Day, J.J., Roitman, M.F., Wightman, R.M., Carelli, R.M., 2007. Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. Nat. Neurosci. 10, 1020–1028.

Dayan, P., 1993. Improving Generalisation for Temporal Difference Learning: The Successor Representation. Neural Comput. 5, 613–624.

Dayan, P., 2012. How to set the switches on this thing. Curr. Opin. Neurobiol. 22, 1068–1074.

de Gardelle, V., Waszczuk, M., Egner, T., Summerfield, C., 2013. Concurrent Repetition Enhancement and Suppression Responses in Extrastriate Visual Cortex. Cereb. Cortex 23, 2235–2244.

Delgado, M.R., Nystrom, L.E., Fissell, C., Noll, D.C., Fiez, J.A., 2000. Tracking the hemodynamic responses to reward and punishment in the striatum. J. Neurophysiol. 84, 3072–3077.

Deshmukh, S.S., Knierim, J.J., 2013. Influence of local objects on hippocampal representations: Landmark vectors and memory. Hippocampus 23, 253–267.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–80.

Diba, K., Buzsáki, G., 2007. Forward and reverse hippocampal place-cell sequences during ripples. Nat. Neurosci. 10, 1241–1242.

Dickinson, A., Balleine, B., 1994. Motivational control of goal-directed action. Anim. Learn. Behav. 22, 1–18.

Dickinson, A., Mulatero, C.W., 1989. Reinforcer specificity of the suppression of instrumental performance on a non-contingent schedule. Behav. Processes 19, 167–180.

Doeller, C.F., Barry, C., Burgess, N., 2010. Evidence for grid cells in a human memory network. Nature 463, 657–61.

Dolan, R.J., Dayan, P., 2013. Goals and habits in the brain. Neuron 80, 312–325.

Dordek, Y., Soudry, D., Meir, R., Derdikman, D., 2016. Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. Elife 5, 1–36.

Dove, A., Pollmann, S., Schubert, T., Wiggins, C.J., Yves von Cramon, D., 2000. Prefrontal cortex activation in task switching: an event-related fMRI study. Cogn. Brain Res. 9, 103–109.

Dragoi, G., Tonegawa, S., 2011. Preplay of future place cell sequences by hippocampal cellular assemblies. Nature 469, 397–401.

Dupret, D., O'Neill, J., Csicsvari, J., 2013. Dynamic reconfiguration of hippocampal interneuron circuits during spatial learning. Neuron 78, 166–80.

Dusek, J.A., Eichenbaum, H., 1997. The hippocampus and memory for orderly stimulus relations. Proc. Natl. Acad. Sci. U. S. A. 94, 7109–14.

Duvelle, É., Grieves, R.M., Hok, V., Poucet, B., Arleo, A., Jeffery, K.J., Save, E., 2019. Insensitivity of place cells to the value of spatial goals in a two-choice flexible navigation task. J. Neurosci. 39, 2522–2541.

Eichenbaum, H., 1999. The hippocampus and mechanisms of declarative memory. Behav. Brain Res. 103, 123–133.

Eichenbaum, H., 2013. Memory on time, Trends in Cognitive Sciences. Elsevier Current Trends.

Eichenbaum, H., 2014. Time cells in the hippocampus: A new dimension for mapping memories. Nat. Rev. Neurosci.

Eichenbaum, H., 2017. The role of the hippocampus in navigation is memory. J. Neurophysiol. 117, 1785–1796.

Eichenbaum, H., Cohen, N.J., 2014. Can We Reconcile the Declarative Memory and Spatial Navigation Views on Hippocampal Function? Neuron.

Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M., Tanila, H., 1999. The Hippocampus, Memory, and Place Cells: Is It Spatial Memory or a Memory Space? Neuron 23, 209–226.

Ekstrom, A.D., Kahana, M.J., Caplan, J.B., Fields, T.A., Isham, E.A., Newman, E.L., Fried, I., 2003. Cellular networks underlying human spatial navigation. Nature 425, 184–188.

Engelhard, B., Finkelstein, J., Cox, J., Fleming, W., Jang, H.J., Ornelas, S., Koay, S.A., Thiberge, S.Y., Daw, N.D., Tank, D.W., Witten, I.B., 2019. Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. Nature.

Estrada, E., Higham, D.J., 2010. Network properties revealed through Matrix Functions 52, 696–714.

Estrada, E., Higham, D.J., Hatano, N., 2009. Communicability betweenness in complex networks. Phys. A Stat. Mech. its Appl. 388, 764–774.

Everitt, B.J., Morris, K.A., O'Brien, A., Robbins, T.W., 1991. The basolateral amygdala-ventral striatal system and conditioned place preference: Further evidence of limbic-striatal interactions underlying reward-related processes. Neuroscience 42, 1–18.

Farovik, A., Place, R.J., McKenzie, S., Porter, B., Munro, C.E., Eichenbaum, H., 2015. Orbitofrontal Cortex Encodes Memories within Value-Based Schemas and Represents Contexts That Guide Memory Retrieval. J. Neurosci. 35, 8333–8344.

Fellows, L.K., Farah, M.J., 2007. The Role of Ventromedial Prefrontal Cortex in Decision Making: Judgment under Uncertainty or Judgment Per Se? Cereb. Cortex 17, 2669–2674.

Ferbinteanu, J., Shapiro, M.L., 2003. Prospective and retrospective memory coding in the hippocampus. Neuron 40, 1227–1239.

Fischl, B., Liu, A., Dale, A.M., 2001. Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. IEEE Trans. Med. Imaging 20, 70–80.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. Neuron 33, 341–355.

Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical Surface-Based Analysis: II: Inflation, Flattening, and a Surface-Based Coordinate System. Neuroimage 9, 195–207.

Foster, D.J., 2017. Replay Comes of Age. Annu. Rev. Neurosci. 40, 581–602.

Foster, D.J., Wilson, M.A., 2006. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. Nature 440, 680–683.

Freyja Ólafsdóttir, H., Barry, C., Saleem, A.B., Hassabis, D., Spiers, H.J., 2015. Hippocampal place cells construct reward related sequences through unexplored space. Elife 4, 1–17.

Fyhn, M., Hafting, T., Treves, A., Moser, M.-B., Moser, E.I., 2007. Hippocampal remapping and grid realignment in entorhinal cortex. Nature 446, 190–194.

Garris, P.A., Collins, L.B., Jones, S.R., Wightman, R.M., 2006. Evoked Extracellular Dopamine In Vivo in the Medial Prefrontal Cortex. J. Neurochem. 61, 637–647.

Garvert, M.M., Dolan, R.J., Behrens, T.E.J., 2017. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. Elife 6, 1–20.

Gauthier, J.L., Tank, D.W., 2018. A Dedicated Population for Reward Coding in the Hippocampus. Neuron 99, 179-193.e7.

Gershman, S.J., Niv, Y., 2010. Learning latent structure: Carving nature at its joints. Curr. Opin. Neurobiol.

Gershman, S.J., Norman, K.A., Niv, Y., 2015. Discovering latent causes in reinforcement learning. Curr. Opin. Behav. Sci. 5, 43–50.

Gilboa, A., Marlatte, H., 2017. Neurobiology of Schemas and Schema-Mediated Memory. Trends Cogn. Sci. 21, 618–631.

Gilboa, A., Sekeres, M., Moscovitch, M., Winocur, G., 2014. Higher-Order Conditioning Is Impaired by Hippocampal Lesions. Curr. Biol. 24, 2202–2207.

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

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., Smith, S.M., Van Essen, D.C., 2016. A multi-modal parcellation of human cerebral cortex. Nature

536, 171–178.

Gremel, C.M., Costa, R.M., 2013. Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. Nat. Commun. 4, 2264.

Grieves, R.M., Jeffery, K.J., 2017. The representation of space in the brain. Behav. Processes.

Gupta, A.S., van der Meer, M.A.A., Touretzky, D.S., Redish, A.D., 2010. Hippocampal Replay Is Not a Simple Function of Experience. Neuron 65, 695–705.

Gustafson, N.J., Daw, N.D., 2011. Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. PLoS Comput. Biol. 7, e1002235.

Haber, S.N., Behrens, T.E.J., 2014. The Neural Network Underlying Incentive-Based Learning: Implications for Interpreting Circuit Disruptions in Psychiatric Disorders. Neuron 83, 1019–1039.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., Moser, E.I., 2005. Microstructure of a spatial map in the entorhinal cortex. Nature 436, 801–806.

Hampton, A.N., Bossaerts, P., O'Doherty, J.P., 2006. The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans. J. Neurosci. 26, 8360–8367.

Hare, T.A., O'Doherty, J.P., Camerer, C.F., Schultz, W., Rangel, A., 2008. Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. J. Neurosci. 28, 5623–30.

Harlow, H.F., 1949. The formation of learning sets. Psychol. Rev.

Harlow, J., 1868. Recovery from the passage of an iron bar through the head. Publ. Massachusetts Med. Soc.

Hassabis, D., Kumaran, D., Vann, S.D., Maguire, E.A., 2007. Patients with hippocampal amnesia cannot imagine new experiences. Proc. Natl. Acad. Sci. U. S. A. 104, 1726–1731.

Ho, J., Tumkaya, T., Aryal, S., Choi, H., Claridge-Chang, A., 2019. Moving beyond P values: data analysis with estimation graphics. Nat. Methods 16, 565–566.

Hoffeld, D.R., Kendall, S.B., Thompson, R.F., Brogden, W.J., 1960. Effect of amount of preconditioning training upon the magnitude of sensory preconditioning. J. Exp. Psychol. 59, 198–204.

Hoffeld, D.R., Thompson, R.F., Brogden, W.J., 1958. Effect of stimuli time relations during preconditioning training upon the magnitude of sensory preconditioning. J. Exp. Psychol. 56, 437–442.

Hollup, S.A., Molden, S., Donnett, J.G., Moser, M.B., Moser, E.I., 2001. Accumulation of hippocampal place fields at the goal location in an annular watermaze task. J. Neurosci. 21, 1635–44.

Howard, J.D., Kahnt, T., 2017. Identity-Specific Reward Representations in Orbitofrontal Cortex Are Modulated by Selective Devaluation. J. Neurosci. 37, 2627–2638.

Howe, M.W., Tierney, P.L., Sandberg, S.G., Phillips, P.E.M., Graybiel, A.M., 2013.

Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature 500, 575–579.

Høydal, Ø.A., Skytøen, E.R., Andersson, S.O., Moser, M.-B., Moser, E.I., 2019. Object-vector coding in the medial entorhinal cortex. Nature 568, 400–404.

Høydal, Ø.A., Skytøen, E.R., Moser, M.-B., Moser, E.I., 2018. Object-vector coding in the medial entorhinal cortex. bioRxiv.

Hu, D., Amsel, A., 1995. A simple test of the vicarious trial-and-error hypothesis of hippocampal function. Proc. Natl. Acad. Sci. U. S. A. 92, 5506–9.

Hull, C.L., 1943. Principles of behavior. Appleton-century-crofts New York.

Hunt, L.T., Kolling, N., Soltani, A., Woolrich, M.W., Rushworth, M.F.S., Behrens, T.E.J., 2012. Mechanisms underlying cortical activity during value-guided choice. Nat. Neurosci. 15, 470–476.

Hunt, L.T., Malalasekera, W.M.N., de Berker, A.O., Miranda, B., Farmer, S.F., Behrens, T.E.J., Kennerley, S.W., 2018. Triple dissociation of attention and decision computations across prefrontal cortex. Nat. Neurosci. 21, 1471–1481.

Huys, Q.J.M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S.J., Dayan, P., Roiser, J.P., 2015. Interplay of approximate planning strategies. Proc. Natl. Acad. Sci. 112, 3098–3103.

Izquierdo, A., 2017. Functional heterogeneity within rat orbitofrontal cortex in reward learning and decision making. J. Neurosci. 37, 10529–10540.

Izquierdo, A., Suda, R.K., Murray, E. a, 2004. Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. J. Neurosci. 24, 7540–7548.

Jacobs, J., Weidemann, C.T., Miller, J.F., Solway, A., Burke, J.F., Wei, X.-X.X., Suthana, N., Sperling, M.R., Sharan, A.D., Fried, I., Kahana, M.J., 2013. Direct recordings of grid-like neuronal activity in human spatial navigation. Nat. Neurosci. 16, 1188–1190.

Jadhav, S.P., Kemere, C., German, P.W., Frank, L.M., 2012. Awake hippocampal sharp-wave ripples support spatial memory. Science (80-.). 336, 1454–1458.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. Neuroimage 17, 825–841.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL. Neuroimage 62, 782–790.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5, 143–156.

Jezek, K., Henriksen, E.J., Treves, A., Moser, E.I., Moser, M.-B., 2011. Theta-paced flickering between place-cell maps in the hippocampus. Nature 478, 246–9.

Jocham, G., Brodersen, K.H., Constantinescu, A.O., Kahn, M.C., Ianni, A.M., Walton, M.E., Rushworth, M.F.S., Behrens, T.E.J., 2016. Reward-Guided Learning with and without Causal Attribution. Neuron 90, 177–190.

Joel, D., Niv, Y., Ruppin, E., 2002. Actor–critic models of the basal ganglia: new anatomical and computational perspectives. Neural Networks 15, 535–547.

Johnson, A., Redish, A.D., 2007. Neural Ensembles in CA3 Transiently Encode Paths Forward of the Animal at a Decision Point. J. Neurosci. 27, 12176–12189.

Jones, J.L., Esber, G.R., Mcdannald, M.A., Gruber, A.J., Hernandez, A., Mirenzi, A., Schoenbaum, G., 2012. Orbitofrontal Cortex Supports Behavior and Learning Using Inferred But Not Cached Values. Science (80-.). 338, 953–956.

Juechems, K., Balaguer, J., Herce Castañón, S., Ruz, M., O'Reilly, J.X., Summerfield, C., 2019. A Network for Computing Value Equilibrium in the Human Medial Prefrontal Cortex. Neuron 101, 977-987.e3.

Juechems, K., Balaguer, J., Ruz, M., Summerfield, C., 2017. Ventromedial Prefrontal Cortex Encodes a Latent Estimate of Cumulative Reward. Neuron 93, 705-714.e4.

Juechems, K., Summerfield, C., 2019. Where Does Value Come From? Trends Cogn. Sci.

Julian, J.B., Keinath, A.T., Frazzetta, G., Epstein, R.A., 2018. Human entorhinal cortex represents visual space using a boundary-anchored grid. Nat. Neurosci. 21, 191–194.

Jung, M.W., Wiener, S.I., McNaughton, B.L., 1994. Comparison of Spatial Firing Characteristics of Units in Dorsal and Ventral Hippocampus of the Rat. J. Neurosci. 14, 7347–7356.

Kamin, L., 1969. Selective association and conditioning. Fundam. issues Instrum. Learn.

Karlsson, M.P., Frank, L.M., 2009. Awake replay of remote experiences in the hippocampus. Nat. Neurosci. 12, 913–918.

Kemp, C., Tenenbaum, J.B., 2008. The discovery of structural form. Proc. Natl. Acad. Sci. U. S. A. 105, 10687–92.

Kemp, C., Tenenbaum, J.B., 2009. Structured statistical models of inductive reasoning. Psychol. Rev. 116, 20–58.

Keramati, M., Smittenaar, P., Dolan, R.J., Dayan, P., 2016. Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. Proc. Natl. Acad. Sci. 113, 12868–12873.

Killcross, S., Coutureau, E., 2003. Coordination of Actions and Habits in the Medial Prefrontal Cortex of Rats. Cereb. Cortex 13, 400–408.

Killian, N.J., Buffalo, E.A., 2018. Grid cells map the visual world. Nat. Neurosci. 21, 161–162.

Kimberg, D.Y., Aguirre, G.K., D'Esposito, M., 2000. Modulation of task-related neural activity in task-switching: an fMRI study. Cogn. Brain Res. 10, 189–196.

Klein-Flügge, M.C., Barron, H.C., Brodersen, K.H., Dolan, R.J., Behrens, T.E.J., 2013. Segregated Encoding of Reward-Identity and Stimulus-Reward Associations in Human Orbitofrontal Cortex. J. Neurosci. 33, 3202–3211.

Klein-Flügge, M.C., Hunt, L.T., Bach, D.R., Dolan, R.J., Behrens, T.E.J., 2011. Dissociable Reward and Timing Signals in Human Midbrain and Ventral Striatum. Neuron 72, 654–664.

Knutson, B., Rick, S., Wimmer, G.E., Prelec, D., Loewenstein, G., 2007. Neural

Predictors of Purchases. Neuron 53, 147–156.

Knutson, B., Westdorp, A., Kaiser, E., Hommer, D., 2000. FMRI Visualization of Brain Activity during a Monetary Incentive Delay Task. Neuroimage 12, 20–27.

Köhler, W., 1917. The Mentality of Apes.

Kolling, N., Behrens, T.E.J., Mars, R.B., Rushworth, M.F.S., 2012. Neural mechanisms of foraging. Science 336, 95–8.

Kolling, N., Wittmann, M., Rushworth, M.F.S., 2014. Multiple Neural Mechanisms of Decision Making and Their Competition under Changing Risk Pressure. Neuron 81, 1190–1202.

Komorowski, R.W., Garcia, C.G., Wilson, A., Hattori, S., Howard, M.W., Eichenbaum, H., 2013. Ventral hippocampal neurons are shaped by experience to represent behaviorally relevant contexts. J. Neurosci. 33, 8079–8087.

Koster, R., Chadwick, M.J., Chen, Y., Berron, D., Banino, A., Düzel, E., Hassabis, D., Kumaran, D., 2018. Big-Loop Recurrence within the Hippocampal System Supports Integration of Information across Episodes. Neuron 99, 1342-1354.e6.

Kraus, B.J., Brandon, M.P., Robinson, R.J., Connerney, M.A., Hasselmo, M.E., Eichenbaum, H., 2015. During Running in Place, Grid Cells Integrate Elapsed Time and Distance Run. Neuron 88, 578–589.

Kraus, B.J., Robinson, R.J., White, J.A., Eichenbaum, H., Hasselmo, M.E., 2013. Hippocampal "Time Cells": Time versus Path Integration. Neuron 78, 1090–1101.

Kriegeskorte, N., Goebel, R., Bandettini, P.A., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. 103, 3863–3868.

Kriegeskorte, N., Mur, M., Bandettini, P.A., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 4.

Krupic, J., Burgess, N., O'Keefe, J., 2012. Neural Representations of Location Composed of Spatially Periodic Bands. Science (80-.). 337, 853–857.

Kumaran, D., McClelland, J.L., 2012. Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. Psychol. Rev. 119, 573–616.

Kumaran, D., Summerfield, J.J., Hassabis, D., Maguire, E.A., 2009. Tracking the Emergence of Conceptual Knowledge during Human Decision Making. Neuron 63, 889–901.

Kurth-Nelson, Z., Economides, M., Dolan, R.J., Dayan, P., 2016. Fast Sequences of Non-spatial State Representations in Humans. Neuron 91, 194–204.

Langdon, A.J., Sharpe, M.J., Schoenbaum, G., Niv, Y., 2018. Model-based predictions for dopamine. Curr. Opin. Neurobiol. 49, 1–7.

Langdon, A.J., Song, M., Niv, Y., 2019. Uncovering the 'state': Tracing the hidden state representations that structure learning and decision-making. Behav. Processes 167, 103891.

Langston, R.F., Ainge, J.A., Couey, J.J., Canto, C.B., Bjerknes, T.L., Witter, M.P., Moser, E.I., Moser, M.-B., 2010. Development of the spatial representation

system in the rat. Science 328, 1576–80.

Lara, A.H., Kennerley, S.W., Wallis, J.D., 2009. Encoding of gustatory working memory by orbitofrontal neurons. J. Neurosci. 29, 765–74.

Lavoie, A.M., Mizumori, S.J.Y., 1994. Spatial, movement- and reward-sensitive discharge by medial ventral striatum neurons of rats. Brain Res. 638, 157–168.

Leutgeb, J.K., Leutgeb, S., Treves, A., Meyer, R., Barnes, C.A., McNaughton, B.L., Moser, M.B., Moser, E.I., 2005. Progressive transformation of hippocampal neuronal representations in "morphed" environments. Neuron 48, 345–358.

Lever, C., Burton, S., Jeewajee, A., O'Keefe, J., Burgess, N., 2009. Boundary Vector Cells in the Subiculum of the Hippocampal Formation. J. Neurosci. 29, 9771–9777.

Levy, D.J., Glimcher, P.W., 2011. Comparing apples and oranges: using reward-specific and reward-general subjective value representation in the brain. J. Neurosci. 31, 14693–707.

Lieder, F., Shenhav, A., Musslick, S., Griffiths, T.L., 2018. Rational metareasoning and the plasticity of cognitive control. PLOS Comput. Biol. 14, e1006043.

Liu, Y., Dolan, R.J., Kurth-Nelson, Z., Behrens, T.E.J., 2019. Human Replay Spontaneously Reorganizes Experience. Cell.

Logothetis, N.K., Eschenko, O., Murayama, Y., Augath, M., Steudel, T., Evrard, H.C., Besserve, M., Oeltermann, A., 2012. Hippocampal–cortical interaction during periods of subcortical silence. Nature 491, 547–553.

Lohrenz, T., McCabe, K., Camerer, C.F., Montague, P.R., 2007. Neural signature of fictive learning signals in a sequential investment task. Proc. Natl. Acad. Sci. U. S. A. 104, 9493–8.

Ludvig, E.A., Sutton, R.S., Kehoe, E.J., 2012. Evaluating the TD model of classical conditioning. Learn. Behav. 40, 305–319.

Mackey, S., Petrides, M., 2010. Quantitative demonstration of comparable architectonic areas within the ventromedial and lateral orbital frontal cortex in the human and the macaque monkey brains. Eur. J. Neurosci. 32, 1940–1950.

Mahadevan, S., 2005. Proto-value functions: Developmental reinforcement learning. In: ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning. ACM Press, New York, New York, USA, pp. 553–560.

Manns, J.R., Eichenbaum, H., 2006. Evolution of declarative memory. Hippocampus.

Mark, S., Moran, R., Parr, T., Kennerley, S., Behrens, T., 2019. Transferring structural knowledge across cognitive maps in humans and models. bioRxiv 860478.

Mathis, A., Stemmler, M.B., Herz, A.V.M., 2014. Probable nature of higher-dimensional symmetries underlying mammalian grid-cell activity patterns. Elife 4, 1–29.

McClelland, J.L., McNaughton, B.L., O'Reilly, R.C., 1995. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psychol. Rev. 102, 419–457.

Mcdannald, M. a., Takahashi, Y.K., Lopatina, N., Pietras, B.W., Jones, J.L., Schoenbaum, G., 2012. Model-based learning and the contribution of the orbitofrontal cortex to the model-free world. Eur. J. Neurosci. 35, 991–996.

McDannald, M.A., Lucantonio, F., Burke, K.A., Niv, Y., Schoenbaum, G., 2011. Ventral Striatum and Orbitofrontal Cortex Are Both Required for Model-Based, But Not Model-Free, Reinforcement Learning. J. Neurosci. 31, 2700–2705.

Mckenzie, S., Frank, A.J.J., Kinsky, N.R.R., Porter, B., Rivière, P.D., Eichenbaum, H., Rivie, P.D., 2014. Hippocampal Representation of Related and Opposing Memories Develop within Distinct , Hierarchically Organized Neural Schemas. Neuron 83, 202–215.

Mehta, M.R., Quirk, M.C., Wilson, M.A., 2000. Experience-Dependent Asymmetric Shape of Hippocampal Receptive Fields. Neuron 25, 707–715.

Miao, C., Cao, Q., Ito, H.T., Yamahachi, H., Witter, M.P., Moser, M.B., Moser, E.I., 2015. Hippocampal Remapping after Partial Inactivation of the Medial Entorhinal Cortex. Neuron 88, 590–603.

Miller, K.J., Botvinick, M.M., Brody, C.D., 2017. Dorsal hippocampus contributes to model-based planning. Nat. Neurosci. 20.

Minsky, M., 1954. Neural Nets and the Brain Model Problem. Princeton.

Minsky, M., 1961. Steps toward Artificial Intelligence. Proc. IRE 49, 8–30.

Miranda, B., Malalasekera, W.M.N., Behrens, T.E., Dayan, P., Kennerley, S.W., 2019. Combined model-free and model-sensitive reinforcement learning in non-human primates. bioRxiv 836007.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–33.

Momennejad, I., Russek, E.M., Cheong, J.H., Botvinick, M.M., Daw, N.D., Gershman, S.J., 2016. The successor representation in human reinforcement learning. bioRxiv 1–27.

Momennejad, I., Russek, E.M., Cheong, J.H., Botvinick, M.M., Daw, N.D., Gershman, S.J., 2017. The successor representation in human reinforcement learning. Nat. Hum. Behav. 1, 680–692.

Montague, P.R., Dayan, P., Sejnowski, T.J., 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. 16, 1936–1947.

Muenzinger, K.F., 1938. Vicarious Trial and Error at a Point of Choice: I. A General Survey of its Relation to Learning Efficiency. Pedagog. Semin. J. Genet. Psychol. 53, 75–86.

Muller, N.G., Strumpf, H., Scholz, M., Baier, B., Melloni, L., 2013. Repetition Suppression versus Enhancement--It's Quantity That Matters. Cereb. Cortex 23, 315–322.

Nadel, L., Moscovitch, M., 1997. Memory consolidation, retrograde amnesia and the hippocampal complex. Curr. Opin. Neurobiol. 7, 217–227.

Nau, M., Navarro Schröder, T., Bellmund, J.L.S., Doeller, C.F., 2018. Hexadirectional coding of visual space in human entorhinal cortex. Nat. Neurosci. 21, 188–190.

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: A primer with examples. Hum. Brain Mapp. 15, 1–25.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A Toolbox for Representational Similarity Analysis. PLoS Comput. Biol. 10, e1003553.

Niv, Y., 2019. Learning task-state representations. Nat. Neurosci. 22, 1544–1553.

Noonan, M.A.P., Walton, M.E., Behrens, T.E.J., Sallet, J., Buckley, M.J., Rushworth, M.F.S., 2010. Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. Proc. Natl. Acad. Sci. U. S. A. 107, 20547–52.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., Dolan, R.J., 2004. Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. Science (80-.). 304, 452–454.

O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., Dolan, R.J., 2003. Temporal difference models and reward-related learning in the human brain. Neuron 38, 329–337.

O'Keefe, J., Nadel, L., 1978. The Hippocampus as a Cognitive Map. Oxford: Clarendon Press.

Oja, E., 1982. Simplified neuron model as a principal component analyzer. J. Math. Biol. 15, 267–273.

Ólafsdóttir, H.F., Carpenter, F., Barry, C., 2016. Coordinated grid and place cell replay during rest. Nat. Neurosci. 19, 792–794.

Olton, D.S., 1979. Mazes, maps, and memory. Am. Psychol. 34, 583–596.

Omer, D.B., Maimon, S.R., Las, L., Ulanovsky, N., 2018. Social place-cells in the bat hippocampus. Science 359, 218–224.

Ostlund, S.B., Balleine, B.W., 2008. On habits and addiction: an associative analysis of compulsive drug seeking. Drug Discov. Today Dis. Model. 5, 235–245.

Otto, A Ross, Gershman, S.J., Markman, A.B., Daw, N.D., 2013. The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. Psychol. Sci. 24, 751–61.

Otto, A. R., Raio, C.M., Chiang, A., Phelps, E.A., Daw, N.D., 2013. Working-memory capacity protects model-based learning from stress. Proc. Natl. Acad. Sci. 110, 20941–20946.

Packard, M.G., McGaugh, J.L., 1996. Inactivation of Hippocampus or Caudate Nucleus with Lidocaine Differentially Affects Expression of Place and Response Learning. Neurobiol. Learn. Mem. 65, 65–72.

Padoa-Schioppa, C., 2009. Range-adapting representation of economic value in the orbitofrontal cortex. J. Neurosci. 29, 14004–14.

Padoa-Schioppa, C., 2011. Neurobiology of economic choice: a good-based model. Annu. Rev. Neurosci. 34, 333–359.

Padoa-Schioppa, C., Assad, J. a, 2006. Neurons in the orbitofrontal cortex encode economic value. Nature 441, 223–226.

Padoa-Schioppa, C., Assad, J.A., 2008. The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. Nat. Neurosci. 11, 95–102.

Park, S.A., Miller, D.S., Nili, H., Ranganath, C., Boorman, E.D., 2019. Map making: Constructing, combining, and navigating abstract cognitive maps. bioRxiv 810051.

Parker, N.F., Cameron, C.M., Taliaferro, J.P., Lee, J., Choi, J.Y., Davidson, T.J., Daw, N.D., Witten, I.B., 2016. Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. Nat. Neurosci. 19, 845–854.

Pelletier, G., Fellows, L.K., 2019. A critical role for human ventromedial frontal lobe in value comparison of complex objects based on attribute configuration. J. Neurosci. 39, 4124–4132.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., Frith, C.D., 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 442, 1042–1045.

Peters, J., Büchel, C., 2009. Overlapping and distinct neural systems code for subjective value during intertemporal and risky decision making. J. Neurosci. 29, 15727–34.

Pfeiffer, B.E., Foster, D.J., 2013. Hippocampal place-cell sequences depict future paths to remembered goals. Nature 497, 74–79.

Pickens, C.L., Saddoris, M.P., Gallagher, M., Holland, P.C., 2005. Orbitofrontal lesions impair use of cue-outcome association in a devaluation task 119, 317–322.

Pickens, C.L., Saddoris, M.P., Setlow, B., Gallagher, M., Holland, P.C., Schoenbaum, G., 2003. Different roles for orbitofrontal cortex and basolateral amygdala in a reinforcer devaluation task. J. Neurosci. 23, 11078–11084.

Plassmann, H., O'Doherty, J., Rangel, A., 2007. Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. J. Neurosci. 27, 9984–8.

Plassmann, H., O'Doherty, J.P., Rangel, A., 2010. Appetitive and Aversive Goal Values Are Encoded in the Medial Orbitofrontal Cortex at the Time of Decision Making. J. Neurosci. 30, 10799–10808.

Preston, A.R., Shrager, Y., Dudukovic, N.M., Gabrieli, J.D.E., 2004. Hippocampal contribution to the novel use of relational information in declarative memory. Hippocampus 14, 148–152.

Preuschoff, K., Bossaerts, P., Quartz, S.R., 2006. Neural Differentiation of Expected Reward and Risk in Human Subcortical Structures. Neuron 51, 381–390.

Preuss, T.M., 1995. Do Rats Have Prefrontal Cortex? The Rose-Woolsey-Akert Program Reconsidered. J. Cogn. Neurosci. 7, 1–24.

Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C., Fried, I., 2005. Invariant visual representation by single neurons in the human brain. Nature 435, 1102–7.

Ramnani, N., Elliott, R., Athwal, B.S., Passingham, R.E., 2004. Prediction error for free monetary reward in the human prefrontal cortex. Neuroimage 23, 777–786.

Rescorla, R.A., 1969. Conditioned inhibition of fear resulting from negative CS-US contingencies. J. Comp. Physiol. Psychol. 67, 504–509.

Rescorla, R.A., Cunningham, C.L., 1978. Within-compound flavor associations. J. Exp. Psychol. Anim. Behav. Process. 4, 267–275.

Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Class. Cond. II Curr. Res. Theory 21, 64–99.

Reuter, M., Rosas, H.D., Fischl, B., 2010. Highly accurate inverse consistent registration: A robust approach. Neuroimage 53, 1181–1196.

Reynolds, J.N.J., Hyland, B.I., Wickens, J.R., 2001. A cellular mechanism of reward-related learning. Nature 413, 67–70.

Rich, E.L., Wallis, J.D., 2016. Decoding subjective decisions from orbitofrontal cortex. Nat. Neurosci. 19, 973–980.

Roesch, M.R., Taylor, A.R., Schoenbaum, G., 2006. Encoding of Time-Discounted Rewards in Orbitofrontal Cortex Is Independent of Value Representation. Neuron 51, 509–520.

Rolls, E.T., 1999. Spatial view cells and the representation of place in the primate hippocampus. Hippocampus 9, 467–480.

Rudebeck, P.H., Saunders, R.C., Lundgren, D.A., Murray, E.A., 2017. Specialized Representations of Value in the Orbital and Ventrolateral Prefrontal Cortex: Desirability versus Availability of Outcomes. Neuron 95, 1208-1220.e5.

Rudebeck, P.H., Saunders, R.C., Prescott, A.T., Chau, L.S., Murray, E. a, 2013. Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. Nat. Neurosci. 16, 1140–5.

Rushworth, M.F.S., Hadland, K.A., Paus, T., Sipila, P.K., 2002. Role of the Human Medial Frontal Cortex in Task Switching: A Combined fMRI and TMS Study. J. Neurophysiol. 87, 2577–2592.

Rushworth, M.F.S., Noonan, M.P., Boorman, E.D., Walton, M.E., Behrens, T.E.J., 2011. Frontal Cortex and Reward-Guided Learning and Decision-Making. Neuron 70, 1054–1069.

Sams-Dodd, F., Lipska, B.K., Weinberger, D.R., 1997. Neonatal lesions of the rat ventral hippocampus result in hyperlocomotion and deficits in social behaviour in adulthood. Psychopharmacology (Berl). 132, 303–310.

Sarel, A., Finkelstein, A., Las, L., Ulanovsky, N., 2017. Vectorial representation of spatial goals in the hippocampus of bats. Science 355, 176–180.

Schapiro, A.C., Rogers, T.T., Cordova, N.I., Turk-Browne, N.B., Botvinick, M.M., 2013. Neural representations of events arise from temporal community structure. Nat. Neurosci. 16, 486–92.

Schoenbaum, G., Roesch, M.R., Stalnaker, T.A., Takahashi, Y.K., 2009. A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. Nat. Rev. Neurosci. 10, 885–892.

Schoenbaum, G., Takahashi, Y.K., Liu, T.-L., Mcdannald, M.A., 2011. Does the orbitofrontal cortex signal value? Ann. N. Y. Acad. Sci. 1239, 87–99.

Schuck, N.W., Cai, M.B., Wilson, R.C., Niv, Y., 2016. Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. Neuron 91, 1402–1412.

Schuck, N.W., Niv, Y., 2019. Sequential replay of nonspatial task states in the human hippocampus. Science (80-.). 364, eaaw5181.

Schultz, W., Dayan, P., Montague, P.R., Deichmann, R., Friston, K., Dolan, R.J., 1997. A neural substrate of prediction and reward. Science (80-.). 275, 1593–1599.

Scoville, W.B., Milner, B., 1957. Loss of recent memory after bilateral hippocampal lesions. J. Neurol. Neurosurg. Psychiatry 20, 11–21.

Ségonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. Neuroimage 22, 1060–1075.

Setlow, B., Holland, P.C., Gallagher, M., 2002. Disconnection of the basolateral amygdala complex and nucleus accumbens impairs appetitive Pavlovian second-order conditioned responses. Behav. Neurosci. 116, 267–275.

Seymour, B., O'Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J., Frackowiak, R.S., 2004. Temporal difference models describe higher-order learning in humans. Nature 429, 664–667.

Sharpe, M.J., Chang, C.Y., Liu, M.A., Batchelor, H.M., Mueller, L.E., Jones, J.L., Niv, Y., Schoenbaum, G., 2017. Dopamine transients are sufficient and necessary for acquisition of model-based associations. Nat. Neurosci. 20, 735–742.

Sharpe, M.J., Stalnaker, T.A., Schuck, N.W., Killcross, S., Schoenbaum, G., Niv, Y., 2019. An Integrated Model of Action Selection: Distinct Modes of Cortical Control of Striatal Decision Making. Annu. Rev. Psychol. 70, 53–76.

Shohamy, D., Wagner, A.D., 2008. Integrating Memories in the Human Brain: Hippocampal-Midbrain Encoding of Overlapping Events. Neuron 60, 378–389.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D., 2016. Mastering the game of Go with deep neural networks and tree search. Nature 529, 484–9.

Singh, T., Mcdannald, M.A., Haney, R.Z., Cerri, D.H., Schoenbaum, G., 2010. Nucleus Accumbens Core and Shell are Necessary for Reinforcer Devaluation Effects on Pavlovian Conditioned Responding. Front. Integr. Neurosci. 4, 126.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155.

Smith, D. V, Hayden, B.Y., Truong, T.-K., Song, A.W., Platt, M.L., Huettel, S.A., 2010. Distinct value signals in anterior and posterior ventromedial prefrontal cortex. J. Neurosci. 30, 2490–5.

Sokal, M., 1994. Gestalt psychology in America in the 1920s and 1930s. Gestalt Psychol. Orig. Found. Influ. 87--119.

Sommer, T., 2017. The Emergence of Knowledge and How it Supports the Memory for Novel Related Information. Cereb. cortex 27, 1906–1921.

Stachenfeld, K.L., Botvinick, M.M., Gershman, S.J., 2017. The hippocampus as a predictive map. Nat. Neurosci. 20, 1643–1653.

Stalnaker, T.A., Berg, B., Aujla, N., Schoenbaum, G., 2016. Cholinergic Interneurons Use Orbitofrontal Input to Track Beliefs about Current State. J. Neurosci. 36, 6242–57.

Stalnaker, T.A., Cooch, N.K., Schoenbaum, G., 2015. What the orbitofrontal cortex does not do. Nat. Neurosci. 18, 620–627.

Steinberg, E.E., Keiflin, R., Boivin, J.R., Witten, I.B., Deisseroth, K., Janak, P.H., 2013. A causal link between prediction errors, dopamine neurons and learning. Nat. Neurosci. 16, 966–973.

Stemmler, M.B., Mathis, A., Herz, A.V.M., 2015. Connecting multiple spatial scales to decode the population activity of grid cells. Sci. Adv. 1, e1500816.

Stensola, H., Stensola, T., Solstad, T., Frøland, K., Moser, M., Moser, E.I., 2012. The entorhinal grid map is discretized. Nature 492, 72–78.

Strait, C.E., Blanchard, T.C., Hayden, B.Y., 2014. Reward Value Comparison via Mutual Inhibition in Ventromedial Prefrontal Cortex. Neuron 82, 1357–1366.

Strait, C.E., Sleezer, B.J., Hayden, B.Y., 2015. Signatures of Value Comparison in Ventral Striatum Neurons. PLOS Biol. 13, e1002173.

Strange, B.A., Witter, M.P., Lein, E.S., Moser, E.I., 2014. Functional organization of the hippocampal longitudinal axis. Nat. Rev. Neurosci. 15, 655–669.

Suarez, J.A., Howard, J.D., Schoenbaum, G., Kahnt, T., 2019. Sensory prediction errors in the human midbrain signal identity violations independent of perceptual distance. Elife 8.

Sun, C., Yang, W., Martin, J., Tonegawa, S., 2020. A Hippocampal Code for Experience Via Its Events as Abstract Units. Cell (under Rev.

Sutton, R.S., 1988. Learning to predict by the methods of temporal differences. Mach. Learn. 3, 9–44.

Sutton, R.S., 1991. Dyna, an integrated architecture for learning, planning, and reacting. ACM SIGART Bull. 2, 160–163.

Sutton, R.S., Barto, A.G., 1981. Toward a modern theory of adaptive networks: Expectation and prediction. Psychol. Rev. 88, 135–170.

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning : an introduction. MIT Press.

Syed, E.C.J., Grima, L.L., Magill, P.J., Bogacz, R., Brown, P., Walton, M.E., 2016. Action initiation shapes mesolimbic dopamine encoding of future rewards. Nat. Neurosci. 19, 34–36.

Takahashi, Y.K., Batchelor, H.M., Liu, B., Khanna, A., Morales, M., Schoenbaum, G., 2017. Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards. Neuron 95, 1395-1405.e3.

Takahashi, Y.K., Roesch, M.R., Wilson, R.C., Toreson, K., O'Donnell, P., Niv, Y., Schoenbaum, G., 2011. Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. Nat. Neurosci. 14, 1590–1597.

Taube, J.S., Muller, R.U., Ranck, J.B., 1990. Head-direction cells recorded from the

postsubiculum in freely moving rats. I. Description and quantitative analysis. J. Neurosci. 10, 420–35.

Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.D., 2011. How to Grow a Mind: Statistics, Structure, and Abstraction. Science 331, 1279–1285.

Thorn, C.A., Atallah, H., Howe, M., Graybiel, A.M., 2010. Differential Dynamics of Activity Changes in Dorsolateral and Dorsomedial Striatal Loops during Learning. Neuron 66, 781–795.

Thorndike, E.L., 1898. Animal Intelligence: An Experimental Study of the Associative Processes in Animals, Psychological Review, Monograph Supplements. Macmillan, New York, NY, US:

Thorndike, E.L., 1905. The elements of psychology. AG Seiler, New York, NY, US:

Thorndike, E.L., 1911. Animal Intelligence: Experimental Studies.

Thorndike, E.L., 1933. A theory of the action of the after-effects of a connection upon it. Psychol. Rev. 40, 434–439.

Tobler, P.N., O'Doherty, J.P., Dolan, R.J., Schultz, W., 2006. Human Neural Learning Depends on Reward Prediction Errors in the Blocking Paradigm. J. Neurophysiol. 95, 301–310.

Tolman, E.C., 1932. Purposive behavior in animals and men. Univ of California Press.

Tolman, E.C., 1939. Prediction of vicarious trial and error by means of the schematic sowbug. Psychol. Rev. 46, 318–336.

Tolman, E.C., 1948. Cognitive maps in rats and men. Psychol. Rev. 55, 189–208.

Tolman, E.C., Honzik, C., 1930. Introduction and removal of reward, and maze performance in rats. Univ. Calif. Publ. Psychol. 4, 257–75.

Tolman, E.C., Ritchie, B.F., Kalish, D., 1946. Studies in spatial learning. I. Orientation and the short-cut. J. Exp. Psychol. 36, 13–24.

Tomov, M.S., Dorfman, H.M., Gershman, S.J., 2018. Neural Computations Underlying Causal Structure Learning. J. Neurosci. 38, 3336–17.

Tse, D., Langston, R.F., Kakeyama, M., Bethus, I., Spooner, P.A., Wood, E.R., Witter, M.P., Morris, R.G.M., 2007. Schemas and Memory Consolidation. Science (80-.). 316, 76–82.

Tse, D., Takeuchi, T., Kakeyama, M., Kajii, Y., Okuno, H., Tohyama, C., Bito, H., Morris, R.G.M., 2011. Schema-Dependent Gene Activation and Memory Encoding in Neocortex. Science (80-.). 333, 891–895.

Ungless, M.A., Magill, P.J., Bolam, J.P., 2004. Uniform Inhibition of Dopamine Neurons in the Ventral Tegmental Area by Aversive Stimuli. Science (80-.). 303, 2040–2042.

van der Meer, M.A., Redish, A.D., 2011. Ventral striatum: a critical look at models of learning and evaluation. Curr. Opin. Neurobiol. 21, 387–392.

van der Meer, M.A.A., Redish, A.D., 2009. Covert expectation-of-reward in rat ventral striatum at decision points. Front. Integr. Neurosci. 3, 1.

Vikbladh, O.M., Meager, M.R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., Burgess, N., Daw, N.D., 2019. Hippocampal Contributions to Model-Based

Planning and Spatial Memory. Neuron 102, 683-693.e4.

Waelti, P., Dickinson, A., Schultz, W., 2001. Dopamine responses comply with basic assumptions of formal learning theory. Nature 412, 43–48.

Wallis, J.D., 2011. Cross-species studies of orbitofrontal cortex and value-based decision-making. Nat. Neurosci. 15, 13–19.

Walton, M.E., Behrens, T.E.J., Buckley, M.J., Rudebeck, P.H., Rushworth, M.F.S., 2010. Separable Learning Systems in the Macaque Brain and the Role of Orbitofrontal Cortex in Contingent Learning. Neuron 65, 927–939.

Walton, M.E., Bouret, S., 2019. What Is the Relationship between Dopamine and Effort? Trends Neurosci. 42, 79–91.

Wang, J.X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J.Z., Hassabis, D., Botvinick, M.M., 2018. Prefrontal cortex as a meta-reinforcement learning system. Nat. Neurosci. 21, 860–868.

Ward-Robinson, J., Hall, G., 1998. Backward Sensory Preconditioning when Reinforcement is Delayed. Q. J. Exp. Psychol. Sect. B 51, 349–362.

Wei, X.X., Prentice, J., Balasubramanian, V., 2015. A principle of economy predicts the functional architecture of grid cells. Elife 4, 1–29.

Whittington, J.C.R., Muller, T.H., Mark, S., Chen, G., Barry, C., Burgess, N., Behrens, T.E.J., 2019. The Tolman-Eichenbaum Machine: Unifying space and relational memory through generalisation in the hippocampal formation. bioRxiv 770495.

Wikenheiser, A.M., Marrero-Garcia, Y., Schoenbaum, G., 2017. Suppression of Ventral Hippocampal Output Impairs Integrated Orbitofrontal Encoding of Task Structure. Neuron 95, 1197-1207.e3.

Wikenheiser, A.M., Schoenbaum, G., 2016. Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. Nat. Rev. Neurosci. 17, 513–523.

Wills, T.J., Cacucci, F., Burgess, N., O'Keefe, J., 2010. Development of the hippocampal cognitive map in preweanling rats. Science 328, 1573–6.

Wilson, M.A., McNaughton, B.L., 1994. Reactivation of hippocampal ensemble memories during sleep. Science (80-.). 265, 676–679.

Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., Niv, Y., 2014. Orbitofrontal cortex as a cognitive map of task space. Neuron 81, 267–278.

Wilson, S.P., Wilson, P.N., 2018. Failure to demonstrate short-cutting in a replication and extension of Tolman et al.'s spatial learning experiment with humans. PLoS One 13, e0208794.

Wimmer, G.E., Daw, N.D., Shohamy, D., 2012. Generalization of value in reinforcement learning by humans. Eur. J. Neurosci. 35, 1092–1104.

Wimmer, G.E., Shohamy, D., 2012. Preference by association: how memory mechanisms in the hippocampus bias decisions. Science 338, 270–3.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. Neuroimage 92, 381–397.

Wise, S.P., 2008. Forward frontal fields: phylogeny and fundamental function. Trends Neurosci. 31, 599–608.

Wood, E.R., Dudchenko, P.A., Eichenbaum, H., 1999. The global record of memory in hippocampal neuronal activity. Nature 397, 613–616.

Wood, E.R., Dudchenko, P.A., Robitsek, R.J., Eichenbaum, H., 2000. Hippocampal Neurons Encode Information about Different Types of Memory Episodes Occurring in the Same Location. Neuron 27, 623–633.

Wunderlich, K., Rangel, A., O'Doherty, J.P., 2010. Economic choices can be made using only stimulus values. Proc. Natl. Acad. Sci. U. S. A. 107, 15005–10.

Xie, J., Padoa-Schioppa, C., 2016. Neuronal remapping and circuit persistence in economic decisions. Nat. Neurosci. 19, 855–861.

Yin, H., Barnet, R.C., Miller, R.R., 1994. Second-Order Conditioning and Pavlovian Conditioned Inhibition: Operational Similarities and Differences. J. Exp. Psychol. Anim. Behav. Process. 20, 419–428.

Yin, H.H., Knowlton, B.J., 2006. The role of the basal ganglia in habit formation. Nat. Rev. Neurosci.

Yin, H.H., Knowlton, B.J., Balleine, B.W., 2004. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur. J. Neurosci. 19, 181–189.

Yin, H.H., Ostlund, S.B., Knowlton, B.J., Balleine, B.W., 2005. The role of the dorsomedial striatum in instrumental conditioning. Eur. J. Neurosci. 22, 513–523.

Yoon, K., Buice, M.A., Barry, C., Hayman, R., Burgess, N., Fiete, I.R., 2013. Specific evidence of low-dimensional continuous attractor dynamics in grid cells. Nat. Neurosci. 16, 1077–1084.

Yoon, K., Lewallen, S., Kinkhabwala, A.A., Tank, D.W., Fiete, I.R., 2016. Grid Cell Responses in 1D Environments Assessed as Slices through a 2D Lattice. Neuron 89, 1086–1099.

Zeithamova, D., Dominick, A.L., Preston, A.R., 2012. Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. Neuron 75, 168–179.

Zhou, J., Gardner, M.P.H., Stalnaker, T.A., Ramus, S.J., Wikenheiser, A.M., Niv, Y., Schoenbaum, G., 2019. Rat orbitofrontal ensemble activity contains a multiplexed but value-invariant representation of task structure in an odor sequence task (supplemental information). Curr. Biol. 507376.

Zola-Morgan, S., Squire, L.R., Amaral, D.G., 1986. Human amnesia and the medial temporal region: enduring memory impairment following a bilateral lesion limited to field CA1 of the hippocampus. J. Neurosci. 6, 2950–67.