

# NEURAL PLASTICITY IN DECISION MAKING AND MEMORY FORMATION

by  
Mona Maria Garvert

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



Wellcome Trust Centre for Neuroimaging  
Institute of Neurology  
University College London

July 2016

# DECLARATION

I, Mona Maria Garvert, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

London, 15.07.2016

Mona Maria Garvert

# ABSTRACT

Goal-directed behaviour is characterized by an ability to make inferences without direct experience. This requires a model of the environment and of ourselves, which is flexibly adjusted in light of new incoming information. This thesis uses representational functional magnetic resonance imaging (fMRI) techniques in combination with computational modelling to investigate (1) whether humans can construct models of other people's preferences and whether this process influences their own value representation, and (2) how statistical relationships between discrete, non-spatial objects are combined into a model of the world.

The first part of the thesis investigates how subjective values are computed in an intertemporal choice paradigm, and how these value computations are updated as a consequence of learning about the preferences of another. Critically, subjects' own preferences shift towards those of the other when learning about their choices, suggesting that subjects incorporate new knowledge about others into a model of their own preferences. The underlying mechanism involves prediction errors, which introduce plasticity into subjects' mPFC value representations, in turn resulting in a shift in subjects' own preferences.

The second part of this thesis investigates how relationships between arbitrary objects are represented in the brain. Relational knowledge is often considered analogous to spatial reasoning, where relationships are encoded in a hippocampal-entorhinal 'cognitive map'. Here, I show that maps can also be extracted from the entorhinal cortex for discrete relationships between arbitrary stimuli, and in the absence of conscious knowledge. The representation of abstract knowledge in map-like structures suggests that inferences do not need to rely on direct experiences but can be computed anew from mapped knowledge.

Together, these studies reveal how world models are represented and updated at the level of neural representations, providing a bridge between representational codes and cognitive computations.

# ACKNOWLEDGEMENTS

I am indebted to many people for their support and encouragement during the last years. First and foremost, I would like to thank Ray Dolan and Tim Behrens for their fantastic supervision of my PhD. I often considered myself the luckiest person at the FIL because I had the honour to learn all about science and life from two such unusual scientists and people. Both have been incredibly inspiring and encouraging, and supported me in very different, but complementary ways. I have also always felt at home in the labs they filled with such great people.

Many people have made the years at the FIL a very happy time. Thanks to Helen for making our joint work so much fun, and for sharing the PhD journey with me - the best PhD buddy one could wish for. Thanks to Robb for great discussions when we shared the best desks in the office, and for being a great friend and mentor. Thanks to Laurence for all the scientific and non-scientific advice and the many shared lunches. Thanks to Marta, Zeb and Michael for generously supervising my rotation projects at the FIL and for providing so much valuable input since. Thanks to everyone in the Dolan lab, the Behrens lab and at the FIL overall, for creating such a supportive, inspiring and collaborative environment. I have learned so much from all of you. There are many more people to thank, but I should specifically mention Marion, Suz, Olga, Lone, Peter, Misun, Erie, Archy, Raphael, Philipp, Giles, Marcos, Eleanor, Jill, Emma and Alexa for making the FIL and Oxford such colourful places. It was great fun and very inspiring to be around you. Special thanks to Elisa for very productive café sessions – and for being such great company on this journey towards a PhD. I would also like to thank the people who provide all the support a scientist could wish for, including Marcia, Marina, Peter, David, Al, Jan, Elaine, Ric, Chris, Liam and Rachael.

Thanks to the Wellcome Trust for funding my PhD project. I am really grateful for the many doors the programme has opened for me.

Finally, I would also like to thank my family for their never-ending encouragement, and for making all of this possible. I feel incredibly grateful for all the patient support they have provided over those long years that preceded the completion of this thesis.

Last, but not least, I would like to thank Josep for sharing the ups and softening the downs that came with writing this thesis. His never-ending support and belief in me and the many happy moments we have shared have made these years truly unforgettable.

# CONTENTS

<b>1 Introduction .....</b>	<b>17</b>
1.1 Mechanisms of decision making in the human brain .....	18
1.1.1 Value computation in the brain.....	18
1.1.2 Learning about the value of an action or a reward.....	21
1.1.3 Model-free reinforcement learning.....	22
1.1.4 Modelling reinforcement learning.....	23
1.1.5 Dopamine neurons carry a reward prediction error signal .....	24
1.1.6 Prediction errors outside value-based decision making.....	27
1.1.7 Incorporating social preferences of others into personal choice .....	28
1.2 Work in this thesis related to learning-induced plasticity in value computations.....	29
1.3 Model-based learning .....	30
1.4 Neural implementation of a cognitive map.....	32
1.4.1 Anatomy of the hippocampal formation.....	33
1.4.2 Cognitive maps in the hippocampus and the entorhinal cortex .....	35
1.4.3 Hippocampal place cells and entorhinal grid cells.....	40
1.4.4 Sequence coding by theta phase precession.....	41
1.4.5 Map-based influences on behaviour.....	43
1.4.6 Cognitive maps in non-physical abstract space .....	45
1.5 Thesis overview .....	47
<b>2 Methods for Investigating Physiological Brain Activity and Behaviour .....</b>	<b>49</b>
2.1 Introduction.....	50
2.2 Principles of Magnetic Resonance Imaging (MRI) .....	50
2.2.1 Static magnetic field and magnetization.....	50
2.2.2 Application of a radiofrequency pulse .....	51
2.2.3 Field gradient .....	53

2.3	Functional magnetic resonance imaging (fMRI).....	54
2.3.1	The BOLD signal .....	54
2.3.2	Neurophysiological basis of the BOLD signal.....	55
2.3.3	fMRI data analysis .....	56
2.3.4	Slice time correction.....	56
2.3.5	Bias-correction .....	57
2.3.6	Spatial preprocessing.....	57
2.3.7	Statistical analysis .....	58
2.3.8	Model estimation and statistical inference .....	59
2.3.9	Group inferences .....	61
2.4	Tools for indexing neural computations at the meso-scale.....	61
2.4.1	Repetition suppression as a tool for measuring the similarity of neural representations .....	63
2.4.2	Biophysical mechanism underlying repetition suppression.....	65
2.4.3	Comparison of repetition suppression and multivariate pattern analysis methods.....	67
2.5	Hopfield networks as models of associative memory.....	69
<b>3</b>	<b>Learning about the Preferences of Another Alters Subjective Valuation in Delay Discounting .....</b>	<b>73</b>
3.1	Abstract.....	74
3.2	Introduction .....	74
3.3	Methods .....	76
3.3.1	Participants .....	76
3.3.2	Human partner task.....	76
3.3.3	Estimation of discount rates using Bayesian modelling.....	77
3.3.4	Simulation of the other's choices .....	80
3.3.5	Behavioural modelling of a subject's belief about the other's preferences .....	80
3.3.6	Optimization of choice pairs .....	82
3.3.7	Computer partner and visual display conditions.....	84
3.3.8	Discount rate shift analyses.....	85
3.3.9	Two-partner behavioural experiment.....	86
3.4	Results .....	88

3.4.1	Subjects' behaviour can be modelled using a Bayesian learner .....	88
3.4.2	Discount rates are susceptible to social influence .....	90
3.4.3	Discount rate shifts arise out of a simulation of the other's preferences .....	93
3.4.4	Subjective value changes are induced by learning .....	95
3.5	Discussion .....	96

**4 Learning-induced Plasticity in Medial Prefrontal Cortex Predicts Preference Malleability .....99**

4.1	Abstract .....	100
4.2	Introduction .....	100
4.3	Methods .....	102
4.3.1	Subjects .....	102
4.3.2	Task– fMRI study .....	102
4.3.3	Surprise measure .....	104
4.3.4	Scan procedure, fMRI data acquisition and pre-processing .....	104
4.3.5	Physiological noise .....	105
4.3.6	fMRI data analysis .....	105
4.3.7	Mediation analysis .....	107
4.4	Results .....	108
4.4.1	Subjective value changes are induced by learning .....	108
4.4.2	Plasticity between neural representations of self and other .....	110
4.4.3	Plasticity in mPFC predicts discount rate shifts .....	113
4.4.4	Plasticity in mPFC is predicted by surprise coding in the striatum .....	114
4.5	Discussion .....	116
4.6	Supplementary Figures .....	120

**5 A Map of Abstract Relational Knowledge in Human Entorhinal Cortex ..... 125**

5.1	Abstract .....	126
5.2	Introduction .....	126
5.3	Methods .....	128
5.3.1	Subjects .....	128
5.3.2	Stimuli and task .....	128



5.3.3	fMRI data acquisition and pre-processing .....	130
5.3.4	fMRI data analysis .....	131
5.4	Results .....	135
5.5	Discussion .....	139
5.6	Supplementary Figures .....	143
<b>6</b>	<b>Models of Map-Formation in the Hippocampal-Entorhinal System.....</b>	<b>147</b>
6.1	Abstract.....	148
6.2	Introduction .....	148
6.3	Methods .....	151
6.3.1	Hopfield network with pairwise plasticity.....	151
6.3.2	Eigendecomposition of place cell activity.....	153
6.4	Results .....	156
6.4.1	A memory pattern can be retrieved from a partial cue in an auto-associative Hopfield network.....	156
6.4.2	Pairwise plasticity between neighbouring patterns leads to distance-dependent scaling of representational similarity .....	157
6.4.3	An eigendecomposition of the resulting activity patterns reveals grid cell-like activity.....	161
6.5	Discussion .....	164
6.6	Supplementary Figures .....	169
<b>7</b>	<b>General Discussion.....</b>	<b>171</b>
7.1	Aim .....	172
7.2	Updating models of the world in social decision making.....	172
7.3	Representing the structure of the world .....	175
7.4	Learning-induced acquisition and updating of world models.....	178
7.5	Conclusion.....	179
	<b>References .....</b>	<b>181</b>

# LIST OF FIGURES

1.1	Subjective value coding across the human brain.....	19
1.2	Dopamine neurons encode reward prediction errors.....	25
1.3	Tolman´s setup for exploring cognitive maps in rats.....	31
1.4	Hippocampal anatomy and connectivity between subfields of the hippocampal formation.....	34
1.5	Illustration of place cell activity during spatial navigation.....	36
1.6	Firing fields of a grid cell in the entorhinal cortex.....	38
1.7	Examples of medial entorhinal border cells and head direction cells.....	39
1.8	Hippocampal time cells.....	40
1.9	Hippocampal phase precession.....	42
1.10	Forward-projecting neural representation at a choice point.....	45
2.1	Hydrogen atoms align with the static magnetic field and precess about the z-axis.....	51
2.2	The consequences of applying a radiofrequency pulse to a $B_0$ field.....	52
2.3	Determinants of the BOLD signal.....	54
2.4	The hemodynamic response function.....	55
2.5	Convolution of condition regressors (top) with the HRF (middle) to predict neural activity (bottom).....	60
2.6	Illustration of the principle underlying fMRI adaptation.....	65
2.7	Orientation-selective cortical columns in the primary visual cortex.....	68
2.8	Schematic of a complete, undirected Hopfield network.....	70
2.9	Illustration of a one-dimensional energy surface.....	72
3.1	Experimental design.....	77
3.2	Results of debriefing questionnaire.....	78

---

3.3	Schematic visualization of discount rate estimation.....	79
3.4	Validation of adaptive method. ....	83
3.5	Design of computer partner and visual display control experiments. ....	84
3.6	Discount rate shift binned according to the distance between $k_{self}$ and $k_{other}$ .....	86
3.7	Experimental design of the two-partner version of the experiment.....	87
3.8	Estimation of belief about the other's discount rate in one example subject. ....	89
3.9	Learning the discount rate of another and the consequences for behaviour. ....	90
3.10	Visualization of choice behaviour in one representative subject in the human partner group (top), the computer partner group (middle) and the visual display group (bottom). ....	91
3.11	Relationship between behavioural shift in preference and performance on 'other' trials. ....	92
3.12	Relationship between belief about the other's discount rate, and subjects' shift in preference.....	94
3.13	The shift in preference is induced by learning about the other's preferences. ....	96
4.1	Experimental design of the fMRI experiment. ....	103
4.2	The shift in preference is induced by learning about the other's preferences. ....	109
4.3	Repetition suppression as an index of representational similarity.....	111
4.4	Learning-induced plasticity in mPFC. ....	112
4.5	Relationship between [SN-SF] <sub>1-3</sub> plasticity and shift in discount rate.....	113
4.6	Surprise as a mechanism underlying mPFC plasticity.....	115
4.7	Mediation path diagram for discount rate shift as predicted from a striatal surprise signal.....	115
S4.1	MPFC activity for self, other and value. ....	120
S4.2	Repetition suppression in visual areas.....	121
S4.3	Response time analyses.....	122
S4.4	Relationship between [FN-SN] <sub>1-3</sub> plasticity and shift in discount rate. ....	123
S4.5	Mediation path diagram for discount rate shift as predicted from the mPFC plasticity signal. ....	124
5.1	Experimental design.....	129

---

5.2	fMRI adaptation in the entorhinal cortex decreases with distance on the graph. ....	136
5.3	Relational information is organized as a map. ....	138
S5.1	Task performance. ....	143
S5.2	Anatomically defined regions of interest used for small-volume correction. ....	144
S5.3	Distance-dependent scaling of neural activity is specific to the entorhinal cortex. ....	145
6.1	Schematic depiction of the architecture of the complete, undirected Hopfield network. ....	151
6.2	Simulated place fields. ....	154
6.3	The network retrieves a memory from a partial cue. ....	157
6.4	Plasticity between neighbouring stimuli on the graph induces correlations between network states. ....	158
6.5	Euclidian distances between memories on the graph predict variance over and above the shortest path between memories. ....	159
6.6	Distance-dependent scaling of representational similarity in a large associative graph structure. ....	161
6.7	Eigendecomposition of the task structure. ....	162
6.8	Distance-dependence of correlation between activity vectors across the first 4 principal components. ....	164
S6.1	Network dynamics. ....	169
S6.2	Eigendecomposition of place cell activity for small place fields. ....	170
7.1	Schematic representation of the phenomenon underlying an increase in repetition suppression with learning. ....	174

# LIST OF TABLES

2.1	Image contrast depending on repetition time (TR) and echo time (TE). .....	53
-----	--	----

# ABBREVIATIONS

ACC	Anterior cingulate cortex
ANOVA	Analysis of variance
ATP	Adenosine triphosphate
BOLD	Blood oxygen level dependent
CA	Cornu ammonis (Hippocampal subfields)
CS	Conditioned stimulus
CSF	Cerebral spinal fluid
EEG	Electroencephalography
EPI	Echo planar imaging
FEF	Frontal eye field
fMRI	Functional magnetic resonance imaging
FWE	Family-wise error
FWHM	Full-width at half-maximum
GLM	General linear model
HRF	Hemodynamic response function
iid	Independent and identically distributed
LFP	Local field potential
LTP	Long term potentiation
MDS	Multi-dimensional scaling
MEC	Medial entorhinal cortex
MEG	Magnetoencephalography
mPFC	Medial prefrontal cortex
MNI	Montreal Neurological Institute
MRI	Magnetic resonance imaging
MTL	Medial temporal lobe
MVPA	Multivariate pattern analysis
OFC	Orbitofrontal cortex
PCA	Principal component analysis
ROI	Region of interest
RSA	Representational similarity analysis
RT	Response time

SPM	Statistical parametric mapping
STS	Superior temporal sulcus
SWS	Slow wave sleep
TE	Echo time
TR	Repetition time
TD	Temporal difference
vmPFC	Ventromedial prefrontal cortex
VTA	Ventral tegmental area

---



# 1 INTRODUCTION

Every day we confront countless decisions, ranging from the mundane choice of whether to have ham or cheese on our breakfast toast, or the more consequential decision of whether to become a doctor or a lawyer in 10 years' time. Although not every decision seems so consequential, deciding how to act ultimately influences our chances of survival and our reproductive success. It is therefore of critical importance to efficiently choose between possible actions. The human brain solves this problem by assigning value to the potential courses of action and choosing the action leading to the highest expected reward (Boorman et al., 2009; FitzGerald et al., 2009; Hunt et al., 2012). This strategy requires an adequate representation of the state of our internal and external environment, the putative actions that can be implemented, as well as the outcomes of the different actions. Both the representation of a constantly changing world, and our behavioural policies need to be flexibly adjusted in light of new information or unexpected outcomes in order to maximize reward.

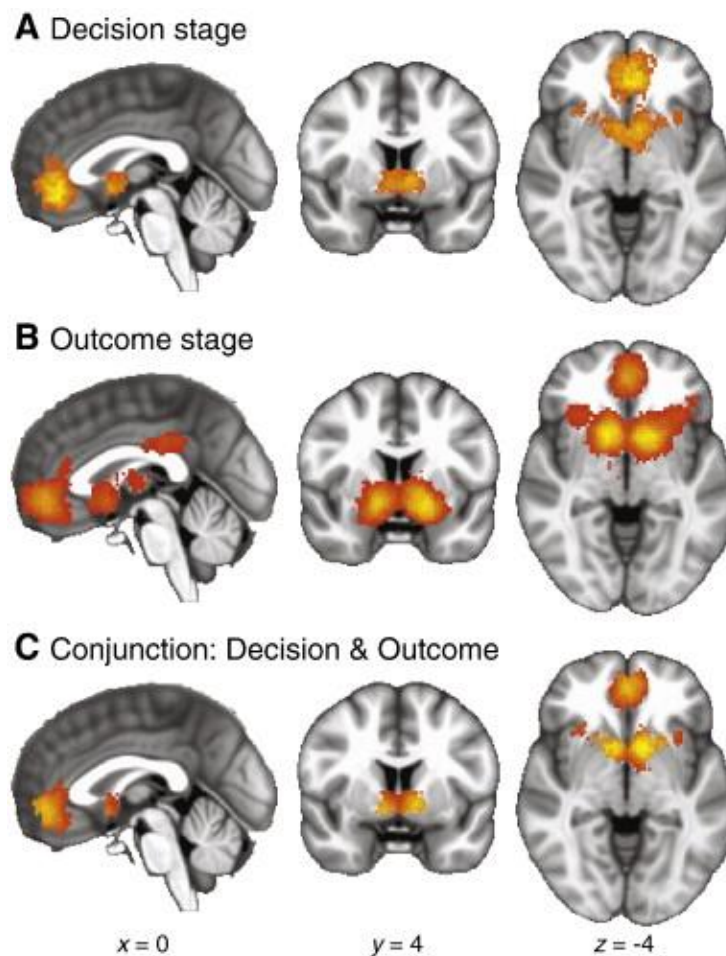
In this thesis, I investigate how the brain learns from experience and organizes new information into a model of the world. In the first part of the thesis, I study how humans compute the value of their choices, and how these value computations are updated as a consequence of learning. Critically, humans do not exclusively learn from their own direct experiences, but can also learn about the environment by observing the behaviour of others. I demonstrate that learning about the preferences of another individual introduces plasticity in subjects' own value computation, which explains a shift in preference. In the second part of the thesis, I address how the brain represents the abstract statistical relationships between elements in the environment to form a model of the world. The representation turns out to be map-like, which is an efficient way of making experiences accessible for goal-directed behaviour, because it allows novel inferences to be made and relationships computed between things that have never been directly experienced before. In both parts of the thesis, I use functional magnetic resonance (fMRI) adaptation to investigate these questions at the level of neural representations.

## **1.1 Mechanisms of decision making in the human brain**

### **1.1.1 Value computation in the brain**

Our brains have evolved to optimize action selection in order to increase our chances of survival. Decision making in the brain is therefore largely guided by our attempts to achieve goals, or receive rewards. A reward is typically defined as anything an animal is ready to work

for, it wants to attain and approach. Avoiding aversive events can also be rewarding. In humans, rewards are often associated with positive subjective states such as pleasure. Rewards include primary reinforcers such as taste, odors, sex, or social affiliation (Rolls, 1999), which have inherent value that does not need to be learned. Other rewards have acquired value through repeated association with primary reinforcers, such as money.



**Figure 1.1 Subjective value coding across the human brain.** Subjective value encoding at the time of evaluating a decision, receiving an outcome or both. The meta-analysis involved data from 206 publications 348 analysis contrasts, 3857 participants, and 3933 activation foci. Reproduced with permission from Bartra et al. (2013).

In the human brain, activity in a wide range of areas tracks the value of an anticipated or actual outcome, including ventromedial PFC (vmPFC), medial orbitofrontal cortex (OFC) and striatum track (Figure 1.1, Hunt et al., 2012; Kable and Glimcher, 2007; Plassmann et al., 2007). Notably, these areas are activated across a wide range of reinforcers, such as expected monetary value (Daw et al., 2006), attractive faces (O'Doherty et al., 2003a) or

preferred drinks (McClure et al., 2004). vmPFC activity also scales positively with the value of a chosen option, and negatively with the value of an unchosen option (Boorman et al., 2009; De Martino et al., 2013).

The prefrontal cortex also integrates information about an animal's current internal motivational state. For example, activity in the orbitofrontal cortex (OFC) correlates with the amount of food only if an animal is hungry (Rolls et al., 1989), and the response in vmPFC incorporates health information in a situation where subjects who successfully control their food intake are tasked to evaluate dietary choices (Hare et al., 2011). Furthermore, mPFC computes values on the fly or without previous experience, for example if information about complex task structures need to be integrated (Hampton et al., 2006), if novel objects need to be evaluated (Barron et al., 2013) or if one's own experience and socially learnt information need to be combined (Behrens et al., 2008).

Patients with damage to their vmPFC often become indecisive even in trivial situations (Barrash et al., 2000) or do not integrate future consequences of their actions in the decision making process (Bechara, 2000; Bechara et al., 1994). A particularly vivid description of the behavioural changes resulting from mPFC damage were provided by Harlow's description of Phineas Gage, who suffered from a dramatic injury to his prefrontal cortex when an iron rod was driven through his head in a railroad accident. His doctor 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)*

Anatomically, mPFC is widely connected to a large array of brain areas, including sensory areas, the hippocampus, the amygdala, the striatum, the insula, the hypothalamus and neuromodulatory systems, such as the dorsal raphe and locus coeruleus (Haber and Behrens, 2014). It can be further subdivided into a lateral OFC part receiving strong sensory inputs,

which is involved in linking stimuli to their subjective value, and a medial OFC part, characterized by strong connections with hypothalamus and visceral control structure in the midbrain. Due to its connectivity to these areas, mPFC has immediate access to a wide range of contextual, sensory and emotional information and can directly influence autonomic and muscular activity. The dorsal mPFC has an overall similar connectivity profile, with stronger connectivity with motor and pre-motor areas and weaker connectivity with emotional and autonomic brain areas, suggesting a critical contribution of this structure to action selection processes (Euston et al., 2012). As a whole, mPFC is thus ideally positioned for assessing the behavioural context and the internal motivational state (Hyman et al., 2012) and mapping events onto an adaptive response to guide behaviour (Euston et al., 2012).

### **1.1.2 Learning about the value of an action or a reward**

In animals, two learning systems are used to aid decision making. A *habitual* or *model-free* system chooses actions based on a reinforcement history and without relying on knowledge about the structure of the world, i.e. actions that were previously rewarded are repeated. While this system is computationally very inexpensive, it is also very inflexible, and habitual behaviours persist when outcomes are devalued. Model-free learning includes *Pavlovian conditioning*, which elicits strong reflexive approach and avoidance behaviours in response to previously neutral stimuli which have been repeatedly paired with a reinforcer. This behaviour is not goal-directed, as it does not influence the chances of obtaining a reward, although there is some evidence that model-based Pavlovian evaluation also exists (Dayan and Berridge, 2014). A *goal-directed* or *model-based* system, on the other hand, explicitly computes the putative outcomes before action selection and implements the choice promising the highest reward. This requires profound knowledge of the contingencies and world relationships in the environment, and the state of the internal and external environment need to be assessed before deciding between alternative actions. The internal model of the world allows behaviour to be flexibly adjusted if an outcome is devalued, but employing a model-based system is computationally very costly. Humans typically display mixed strategies in decision making situations (Loewenstein and O'Donoghue, 2004; Daw et al., 2011).

### 1.1.3 Model-free reinforcement learning

At the beginning of the 20<sup>th</sup> century, behaviourist theories dominated the understanding of behaviour (Ferster and Skinner, 1957; Skinner, 1938; Thorndike, 1898; Watson, 1913). Thorndike, Skinner and others argued that learning goal-directed behaviours occurs when a response is reinforced (stimulus-response learning or S-R learning). S-R learning can be investigated in *operant* (or *instrumental*) *conditioning* paradigms, where rewards are contingent on actions. For example, when trying to escape from a puzzle box, a hungry cat will try a random set of behaviours, until it eventually performs the action that results in a release from the box: a lever press. Thorndike argued that the reward of being released from the box strengthens the association between a stimulus (the puzzle box) and a response (pressing the lever), such that the animal becomes gradually faster at pressing the lever and escaping from the box in subsequent trials. Thorndike named this phenomenon ‘*the law of effect*’ (Thorndike, 1927). Repeating rewarded behaviours can be highly adaptive in stable environments, because it is computationally very inexpensive and S-R learning accounts for the formation of simple reflexes and habits.

*Operant conditioning*, or *S-R learning* contrasts with *classical conditioning*, a phenomenon whereby neutral stimuli acquire value via the repeated association with a reinforcer. Classical conditioning was first investigated by Ivan Pavlov (Pavlov, 1927). He repeatedly paired a neutral stimulus such as the sound of a bell (conditioned stimulus, CS) with a significant event such as the delivery of food (unconditioned stimulus, US). Food typically elicits an innate salivating response in dogs. After a few conditioning trials, the animal started to elicit an anticipatory salivating response when the previously neutral CS was presented, suggesting that the neutral stimulus had acquired value. This simple form of learning from experience can induce Pavlovian approach behaviour if the stimulus is paired with rewards such as food, and it can induce withdrawal behaviour if the stimulus is paired with aversive outcomes such as shocks. An example for such a conditioned response is the approach of a light that predicts the delivery of a reward (Dayan and Balleine, 2002). Because the conditioned response directly results from a prediction of reward, this Pavlovian approach behaviour is so powerful and inflexible that it persists even if it is detrimental to actually receiving a reward. This has been illustrated in a study by Hershberger (1986) who trained hungry chicks that a cup contained food. When he manipulated the cup such that it receded if the animal approached it, and vice versa, the chicks were unable to overcome their Pavlovian approach response and move backwards in order to get to the food. Critically, conditioning occurs only if there

is contingency between the CS and the US, and if no other CS already fully predict the US (Kamin, 1969). This suggests that model-free learning occurs via error-driven learning.

#### 1.1.4 Modelling reinforcement learning

In the 1970s, error-driven prediction learning was first formally described by the Rescorla-Wagner model (Rescorla and Wagner, 1972). The basic principle of the model is that animals learn from prediction errors, and the strength of a CS-US association,  $V_t$ , should be updated relative to the difference between the actual outcome and the predicted outcome. In its simplest form, this learning rule can be described by the following equation:

$$V_{t+1} = V_t + \alpha(r_t - V_t) \quad (2.1)$$

Here, the expected value of the reward in trial  $t+1$  is given by the reward value in trial  $t$  ( $V_t$ ), plus a weighted prediction error term,  $(r_t - V_t)$ , corresponding to the difference between the received reward  $r_t$  and the expected reward  $V_t$ . Changes in expected value are therefore driven by the prediction error, which is maximally positive if a large, unexpected reward is received, 0 if a reward is perfectly predicted and negative if an expected reward is omitted.  $\alpha$  corresponds to a learning rate (with  $0 < \alpha \leq 1$ ), indicating the rate at which associations change in time.

Despite its simplicity, the Rescorla-Wagner model has proven very powerful in explaining a number of learning phenomena, such as reported in Pavlovian conditioning. However, it has two major shortcomings. Firstly, time is discretized in terms of trials, such that the framework does not account for the precise prediction of the time at which a reward can be expected. Secondly, the Rescorla-Wagner fails to explain second order conditioning, where a stimulus A predicts a reward ( $A \rightarrow r$ ) if A predicts B ( $A \rightarrow B$ ), and B predicts reward ( $B \rightarrow r$ ). In reality, second order conditioning is only effective if A precedes B. The Rescorla-Wagner model does not differentiate this situation from a scenario where B precedes A, and thus fails to account for the temporal credit assignment problem.

A more generalized model that overcomes these shortcomings is the *temporal difference learning model*, which assigns a value to any given state in time based on the sum of expected future reward  $V_t$  (Sutton, 1988; Sutton and Barto, 1990). The full algorithm is complex and beyond the scope of this thesis, but it is worth noting that learning of state values occurs via updating through the temporal difference prediction error defined at any point in time  $\delta_{t+1}$ :

$$\delta_{t+1} = r_{t+1} + V_{t+1} - V_t \quad (2.2)$$

where  $r_{t+1}$  is the reward at time  $t+1$ ,  $V_{t+1}$  is the state value at time  $t+1$  and  $V_t$  is the state value at time  $t$ . If an unexpected reward is delivered,  $\delta_{t+1}$  is positive, because the reward value  $r_{t+1}$  is positive, but the change in expectation of future rewards  $V_{t+1} - V_t$  is 0. The delivery of a fully predicted reward results in no prediction error even though  $r_{t+1}$  is positive, as the agent proceeds from a valuable state  $V_t$ , in which a reward was to be expected, to a less valuable state  $V_{t+1}$ , which is no longer predictive of a reward. The unexpected appearance of a CS which predicts the appearance of a reward leads to a positive  $\delta_{t+1}$  despite a reward  $r_{t+1}$  of 0, because the animal transitions to a more valuable state  $V_{t+1}$  that is predictive of a future reward, i.e.  $V_{t+1} > V_t$ . Temporal difference learning is therefore exactly consistent with the behavioural observations subsumed under the term classical conditioning.

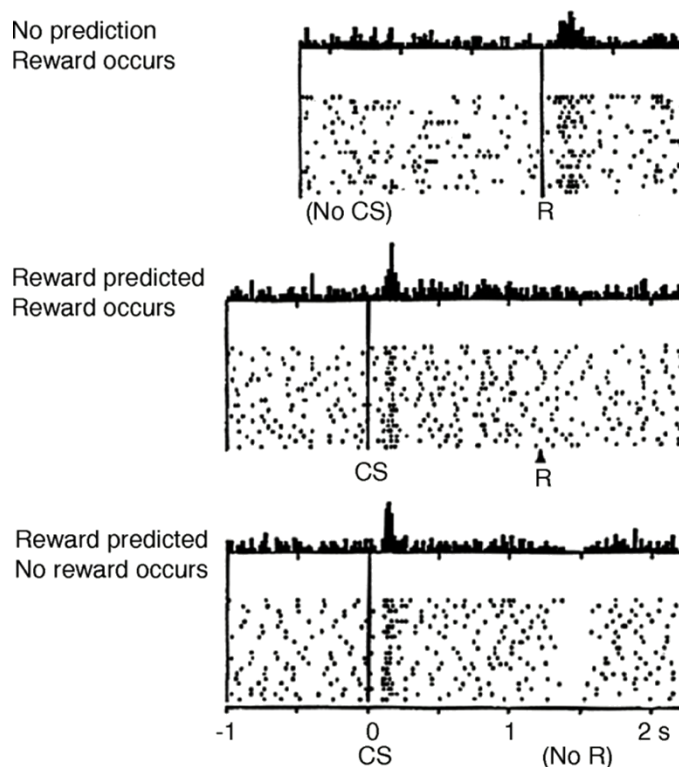
S-R learning is typically modelled using an *actor-critic* architecture, where a value function ('critic') and a policy structure ('actor') are represented independently. Here, the actor selects actions, and the critic evaluates the outcome of the performed actions. As a consequence, the critic learns an evaluation function for each state, reflecting the expected future reward given the typical actions taken from that state. Learning occurs through a temporal difference signal, which drives both learning of the mapping of states onto values ('critic') and of states onto an actions ('actor'). If the TD error is positive and the outcome is better than expected, the mapping of state  $s_t$  onto action  $a_t$  is strengthened. If the TD error is negative, the mapping of state  $s_t$  onto action  $a_t$  is weakened. The actor critic architecture thereby prevents the temporal credit assignment problem. However, as it is based on 'trial-and-error' it is a slow method for learning from the environment, and it does not explain how behaviours that were never reinforced can be flexibly employed.

### 1.1.5 Dopamine neurons carry a reward prediction error signal

In the brain, activity of dopaminergic neurons in the ventral tegmental area (VTA) and substantia nigra pars compacta display responses consistent with the reward prediction error term in the temporal difference learning model (Schultz et al., 1997, Figure 1.2). Before learning, the unexpected presentation of a reward, such as receipt of a drop of fruit juice, leads to a burst of activity in dopamine neurons. If a reward is perfectly predicted because the association between a CS and the reward has been learnt, dopamine neurons respond to the presentation of the CS, but not to reward receipt. If, however, the reward is unexpectedly



omitted after the presentation of the predictive CS, activity of dopamine neurons is suppressed at the time when the reward would have occurred. In line with the predictions of the temporal difference learning model, the size of the dopamine burst scales with the expected value of the reward. The response to the reward itself is maximal if it is unexpected, and decreases with an increase in probability of receiving a reward on a partial reinforcement task; the opposite pattern is observed for the response to the CS (Fiorillo et al., 2003).



**Figure 1.2 Dopamine neurons encode reward prediction errors.** Peri-stimulus time histograms and raster plots from a dopamine neuron in three experimental conditions: no prediction, reward occurs (top), reward predicted, reward occurs (middle) and reward predicted, no reward occurs (bottom). CS, conditioned stimulus, R reward. Reproduced with permission from Schultz et al. (1997).

Signals consistent with prediction errors can also be measured in the human substantia nigra using direct intracranial recordings (Zaghloul et al., 2009) and in the VTA (D’Ardenne et al., 2008; Klein-Flügge et al., 2011) as well as in the human striatum, a region which receives dense inputs from dopaminergic midbrain nuclei (Gläscher et al., 2010; McClure et al., 2003; Niv et al., 2012; O’Doherty et al., 2003b). The dorsal striatum specifically supports stimulus-response learning (Packard and Knowlton, 2002) and has been implicated in carrying prediction errors in instrumental tasks, but not in situations where value prediction errors occur in the absence of action selection (O’Doherty et al., 2004). The ventral striatum, on

the other hand, responds to primary rewards and reward prediction errors more generally (O'Doherty et al., 2004). This anatomical dissociation in expression is consistent with the actor-critic dissociation in reinforcement learning. The signal in the dorsal striatum might reflect the policy improvement function supported by the 'actor', and the signal in the ventral striatum might reflect the state evaluation signal supported by the 'critic'.

The magnitude of a striatal prediction error signal is modulated by drugs that affect dopamine activity (Pessiglione et al., 2006), and prediction error signals in dorsolateral striatum are impaired in Parkinson's patients (Schonberg et al., 2010), who suffer from a loss of dopamine neurons in the substantia nigra. In line with an important role for dopaminergic prediction errors during learning, Parkinson's patients are also impaired in learning from positive outcomes (Rutledge et al., 2009). MRI experiments provide further evidence for a critical contribution of a prediction error signal to learning a stimulus-reward association. For example, those subjects who learned to perform a four-armed bandit task successfully displayed a larger prediction error signal than non-learners (Schönberg et al., 2007).

However, the first causal relationship between the release of dopamine and learning from a stimulus in the environment was provided by a recent optogenetics study. The study capitalized on the phenomenon of blocking, whereby the association of a CS with reward can be prevented if the CS forms a compound together with a second CS which already fully predicts the reward. In this situation, no dopamine is released in response to the new CS, and the animal does not express any behavioural response. If, however, dopamine release is triggered optogenetically when the compound cue is presented, a behavioural response develops to the new CS alone, demonstrating a causal relationship between dopamine release and learning (Steinberg et al., 2013).

VTA neurons putatively exert their effects on learning via projections to the striatum, the hippocampus, the amygdala and neocortical areas such as the medial prefrontal, cingulate and perirhinal cortex (Haber and Behrens, 2014). Dopamine release in these downstream brain areas affects plasticity in cortico-striatal, hippocampal and meso-cortical synapses (Reynolds and Wickens, 2002; Wickens et al., 1996). For example, dopamine released in a novel spatial situation in the hippocampus directly influences plasticity in CA1 (Li et al., 2003). In PFC, direct perfusion with dopamine, the application of a dopamine reuptake inhibitor or the increased release of transient dopamine through stimulation of VTA all lead to a sustained increase in long term potentiation (LTP) (Garris et al., 2006). By signalling the behavioural relevance of an external stimulus or action, dopamine can therefore directly

influence cortical representations. However, the optimal amount of prediction-error induced learning depends on the volatility of the environment as tracked by the anterior cingulate cortex (Behrens et al., 2007).

### 1.1.6 Prediction errors outside value-based decision making

Prediction errors can also be observed outside of reward-guided decision making, for example when sensory expectations (Näätänen et al., 1989; den Ouden et al., 2010) or when stimulus transition probabilities (Meyer and Olson, 2011) are violated. This demonstrates the importance of prediction error signals as a domain-general teaching signal for learning about the environment. This notion has been formalized in hierarchical predictive coding models of brain function (Friston, 2009), initially developed to explain the properties of simple cell receptive fields in the primary visual cortex (Rao and Ballard, 1999). Predictive coding assumes that *top-down* projections from downstream brain areas carry a prediction signal pertaining to the expected neural dynamics, which is constantly compared to *bottom-up* activity at a lower level of the cortical hierarchy. The brain aims to minimize the prediction error signal resulting from the discrepancy between the expected signal and the actual signal by optimizing the connectivity within cell assemblies at those levels of the hierarchy that experience prediction errors (Friston, 2010). Intriguingly, recent MEG studies provide evidence that bottom-up sensory information and a top-down expectancy signals are carried by synchronization in different frequency bands (Bastos et al., 2015; Michalareas et al., 2016).

The prediction error theory of perceptual processing has been tested in a range of fMRI and MEG experiments, for example by manipulating expectancy through varying the frequency at which face stimuli were presented (Summerfield et al., 2008, 2011). Summerfield et al. (2008) observed an attenuated response to stimuli which occurred frequently (75% of trials) compared to stimuli whose occurrence was rare (25% of trials), suggesting that stimulus expectation modulates the size of a perceptual prediction error. Similar expectation suppression effects are observed in auditory oddball paradigms (Garrido et al., 2007), in response to visually presented shapes (Stefanics et al., 2011), and in somatosensory stimulation (Valentini et al., 2011). They are also consistent with heightened blood oxygenation level dependent (BOLD) signal or evoked responses to unexpected or novel stimuli (Egner et al., 2010; Näätänen et al., 1989; Ouden et al., 2009; Strange et al., 2005a). Evidence for prediction error coding can also be found in direct electrophysiological

recordings of monkey IT neurons, which respond more strongly when a presented image violates a learned statistical transition rule (Meyer and Olson, 2011).

### **1.1.7 Incorporating social preferences of others into personal choice**

As humans, we not only learn from our own experiences but we can also derive inferences about the world from observing the behaviour of others. In fact, it has been argued the cultural success of mankind is attributable to our exceptional ability to learn in social contexts, and to cooperate with others (Boyd et al., 2011; Dunbar and Shultz, 2007; Frith and Frith, 2010). Social interactions between humans are highly complex, and in many ways our mental life is co-dependent on that of other human beings. A central element of social interactions is understanding behaviour of others as an outcome of their internal mental states. The ability to infer the mental states of another person, their beliefs and preferences is referred to as Theory of Mind or mentalizing (Premack and Woodruff, 1978).

It has long been a matter of debate whether the computations involved in processing information about social relationships and the mental states of conspecifics require specialized brain systems or whether the mechanisms involved in other cognitive domains also underlie social behaviours. The discovery of mirror neurons, which respond both to the execution and to the observation of an action (Kilner et al., 2009; Rizzolatti et al., 1996, 2001), provided a hint that the same principles that underlie cognitive computations for self might also serve as a cortical substrate for understanding others. In line with this notion, experiencing pain and observing other people's pain activates the anterior cingulate cortex (Jackson et al., 2005; Singer et al., 2004), experiencing and observing disgust in others activates the insula (Wicker et al., 2003) and choosing for self and for others activates medial prefrontal cortex (Jenkins et al., 2008; Nicolle et al., 2012). It has also been demonstrated that the prediction error signals controlling reinforcement learning also guide learning about a confederate player's choices in a game (Behrens et al., 2008). This suggests that the signals that guide learning about stimuli or actions also guide learning about a confederate's intentions.

The network of brain regions involved in attributing mental states to other people includes the dorsomedial prefrontal cortex (dmPFC), the temporo-parietal junction, the precuneus/posterior sulcus, superior temporal sulcus and temporal poles (Amodio and Frith, 2006; Saxe, 2006). Traditionally, it has been believed that these regions perform uniquely social computations. However, a recent experiment challenges this notion and instead

suggests that dmPFC also activity reflects value in situations where choice is made in an abstract frame of reference, which is particularly pertinent in situations where one's own sensory and motor environment needs to be ignored. For example, if subjects choose for themselves in an intertemporal choice paradigm, their own subjective values are represented in vmPFC, while the subjective values of another individual whose preferences were learnt previously are simultaneously represented in dmPFC (Nicolle et al., 2012). Traditionally, this would have been taken to suggest that dmPFC performs social value computations. However, if roles are reversed and subjects execute a choice on behalf of the other individual, the other's preferences are now represented in vmPFC while the subject's own preferences are represented in dmPFC. This suggests that dmPFC instead represents a modelled choice, i.e. the values of the agent not currently relevant for behaviour, whereas a circuitry in vmPFC is involved in executing choice irrespective of an agent's identity.

## **1.2 Work in this thesis related to learning-induced plasticity in value computations**

In this thesis, I have used fMRI adaptation in combination with computational modelling to investigate how learning about another person's preferences induces plasticity in a vmPFC value computation. Due to the shared circuitry computing value for self and other, introducing plasticity in a population computing value for another should lead to a simultaneous update of a value computation for self. This could potentially have profound consequences for one's own preferences, and might thereby explain the social conformity effects observed when learning about other people's opinions or memories (Campbell-Meiklejohn et al., 2010; Edelson et al., 2011; Klucharev et al., 2009; Zaki et al., 2011). Such plasticity is likely driven by one of two types of social prediction errors. Confederate prediction errors can arise as we are updating our beliefs about the preferences of another individual. Such prediction errors could provide a teaching signal about the other's preferences. However, a prediction error could also arise from observing another's behaviour and comparing it to the most likely behaviour we would have executed given the same context. This includes social expectancy prediction errors, signalling the discrepancy between our own preferences, and the choices of a social group (Campbell-Meiklejohn et al., 2010; Harris and Fiske, 2010; Klucharev et al., 2009). I demonstrate that such a 'self-referential' prediction error, experienced when comparing the other's choice to the decision one would

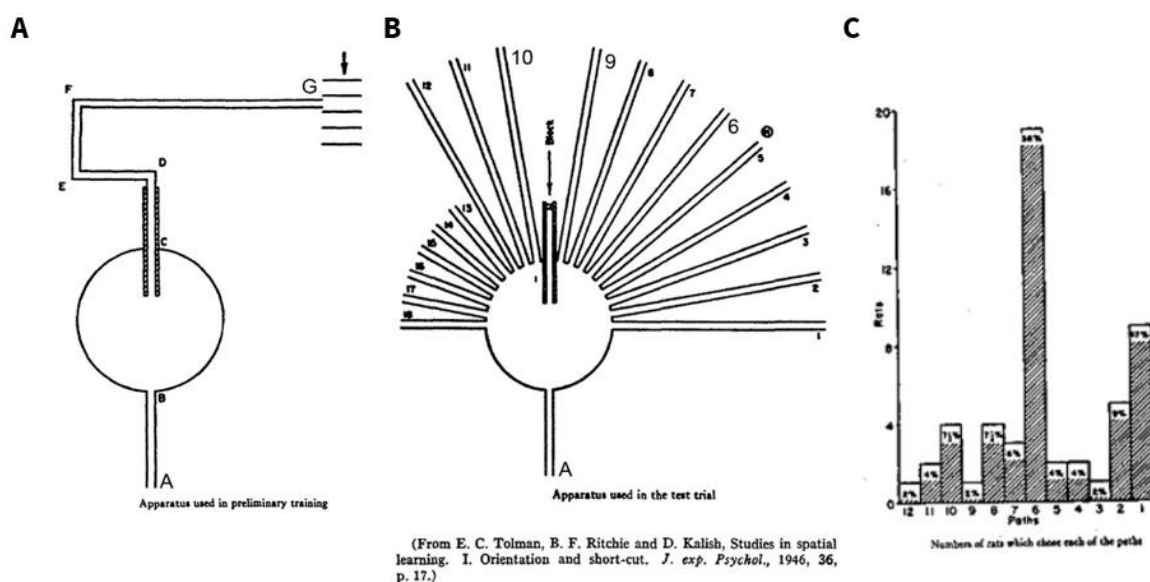
have made in the same situation oneself, predicts a mPFC plasticity effect, whereby the representational similarity between another person's value computation and one's own value computation increases. This plasticity effect in turn predicts a subject's shift in preference, suggesting that social influence can be understood on the level of neural populations, as the consequence of a learning-induced plasticity in mPFC.

### 1.3 Model-based learning

Many aspects of behaviour can be learned through reinforcement of initially random movements, which ultimately leads to the formation of habits. However, relying on this system alone is insufficient in situations where the outcome is delayed, depends on a series of actions ('credit assignment problem'), or requires an internal model that can be flexibly used to update behavioural policies in light of new information.

The existence of such models was first demonstrated by Tolman in the 20<sup>th</sup> century. Tolman introduced the notion of *latent learning*, whereby animals find the location of a food reward in a novel environment faster if they previously have the opportunity to explore the environment (Tolman and Honzik, 1930). He investigated this by placing two groups of rats in a complex T-maze. One group was rewarded whenever they reached the end of the maze from day 1 onwards. Group 2 were not rewarded on the first ten days, and only received rewards from day 11 onwards. If animals simply learn through reinforcement, rats in group 2 should show the same behaviour from day 11 onwards as rats in group 1 show from day 1. Instead, Tolman observed that while the immediately rewarded animals performed better on the first 10 days of the experiment, when food was introduced for group 2 these animals showed a large decrease in response time and errors, and in fact performed better than the animals in group 1. This suggests that rats in group 2 acquired knowledge about the structure of the maze in the first ten days, even though it was not behaviourally relevant or rewarded. When they were later motivated to perform accurately, the acquired knowledge about the structure of the maze allowed the animals to quickly find the correct path to the rewarded location. Tolman called this type of learning *latent learning* and proposed that the animals must be forming a 'cognitive map' of the environment without explicit reinforcement, which allows them to find a path towards a goal when needed. This behaviour cannot be explained by Thorndike's *law of effect*, since no actions were reinforced during the initial exploratory phase.

In a second study, Tolman tested whether rats are able to use the model they have acquired of the world to update their behavioural strategies in light of new incoming information. He trained animals on a maze task, in which they had to find the path from a start location A to a reward location G (Figure 1.3A). After multiple training days, the initial arm of the maze leading to a reward was blocked, and instead multiple radially arranged paths were provided (Figure 1.3B). Tolman observed that a large proportion of animals successfully chose path 6, a shortcut which directly led to the reward location from the start location (Figure 1.3C). Notably, the animals had never experienced this particular path before, so that it could not have been learned from the reinforcement history. Instead, animals would only be able to compute this new shortest path if they have an explicit representation of the relationship between their location, the reward location and the paths between start and goal locations. Tolman suggested that the animal had constructed a ‘cognitive map’ of the environment, which allowed for the relationship between landmarks to be computed flexibly and for shortcuts or detours around obstacles to be planned (Tolman, 1948). Notably, while Tolman deserves to be credited with the notion of a ‘cognitive map’ facilitating flexible, goal-directed decision making, attempts to replicate his experimental results have not always been successful (Gentry et al., 1947; Muir and Taube, 2004).



**Figure 1.3** Tolman’s setup for exploring cognitive maps in rats. **A** Apparatus used in preliminary training. Rats entered the arena in location A. The food reward was located in position G. **B** After multiple training days, the original path was blocked. Instead, animals could choose from 18 radially arranged paths. The direct path leading to the reward was path 6. **C** Number of rats choosing paths 1-12 displayed in B on the test trial. Most rats chose path 6. A second peak could be observed for path 1, which might relate to the fact that the most recent turn before reaching the goal in the original setting was a right turn. Adapted from Tolman (1948).

Unlike the model-free habitual system, this type of model-based system is very flexible as it accounts for changes in the environment or internal state. However, it is computationally very costly to compute the likely outcomes of all possible actions, and in many instances the problem becomes non-tractable. In highly practiced tasks in stable environments it can therefore be advantageous to rely on a computationally efficient, fast and highly trained behavioural policy as provided by the model-free habitual system (Daw et al., 2005). It is now believed that such model-free and model-based systems co-exist in the human brain (Daw et al., 2005) and human behaviour typically shows a mixture of model-free and model-based behaviour (Daw et al., 2011). Their relative influence on behaviour is best seen in devaluation paradigms. For example, hungry rats learn to press a lever in order to obtain food. When food is paired with illness, rats develop an aversion against the food and avoid pressing the lever if they have been moderately trained on the task (Holland). In this situation, the animals are sensitive to the outcome of their action, and realize it is no longer aligned with their goal. If, however, the rats have previously been overtrained on the lever-pressing task, their behaviour becomes insensitive to the fact that the outcome has been devalued and they continue pressing the lever. This suggests that with repetition the action has become automatic, or habitual. Crucially, habitual and value-based behaviours are supported by dissociable neural circuits. Animals who experience lesions to dorsolateral striatum never form habits and remain sensitive to devaluation even after overtraining (Yin et al., 2004). If prefrontal areas or the dorsomedial striatum are lesioned, on the other hand, rats behave habitually and are insensitive to devaluation irrespective of the amount of training (Yin et al., 2005).

## **1.4 Neural implementation of a cognitive map**

Despite the fact that Tolman's experiment specifically investigated maps in spatial navigation, his notion of a cognitive map as an organizational principle went far beyond physical space. He hypothesized that other - or potentially all - types of relational information might be organized in a map-like fashion, which could then be used to guide model-based reasoning. Such organization of knowledge would allow for paths to be computed through an abstract 'concept space' which could underlie flexible model-based behaviour. This also means that the computations and neural codes involved in spatial information processing might be equally relevant for operations performed on the relationship between non-spatial



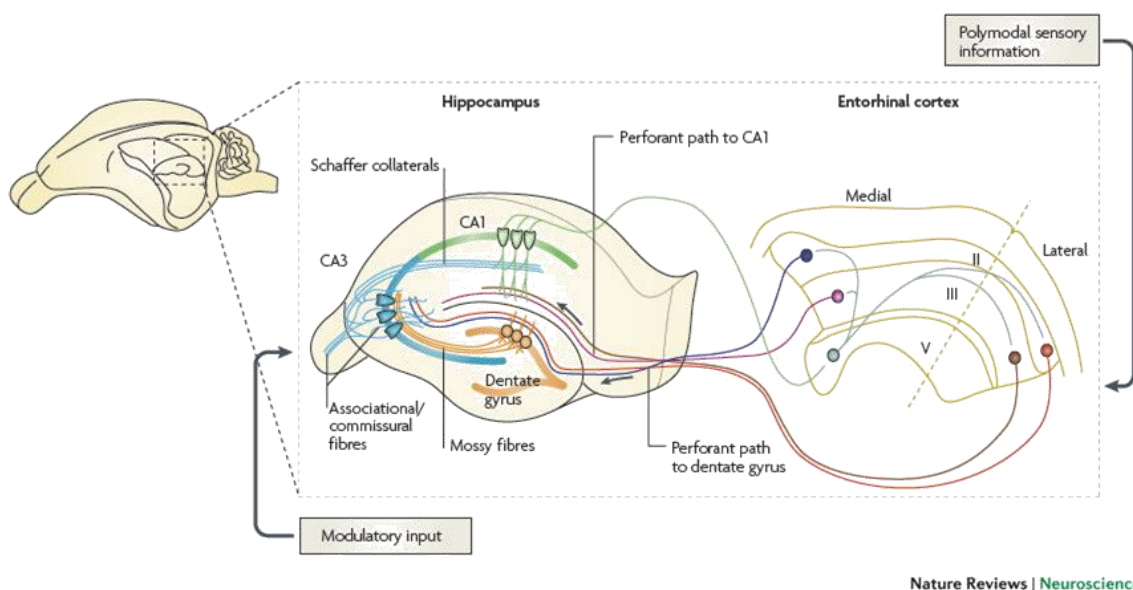
information. However, Tolman did not make explicit assumptions about the neural instantiation of this map in the brain.

In the 1970s, John O'Keefe discovered hippocampal 'place cells' whose firing activity is precisely localized in space (O'Keefe and Dostrovsky, 1971). He hypothesized that these cells could form the physiological and psychological basis of a 'cognitive map' (O'Keefe and Nadel, 1978). A few years earlier, the critical contribution of the hippocampal formation to another, seemingly disparate aspect of cognition, namely episodic memory, had been discovered (Scoville and Milner, 1957) based on studies in one of the most famous neuropsychological patients in history, Henry Molaison (H.M.). H.M. underwent bilateral medial temporal lobe (MTL) resection to treat his temporal lobe epilepsy. After the surgery, H. M. was found to suffer from severe anterograde amnesia, with no ability to encode new episodic or declarative memories. These seemingly separate accounts of the role of hippocampal function in spatial navigation and episodic memory co-existed in parallel for a long time, and converged only recently, when the first evidence appeared that indeed the physiological mechanisms evolved to support spatial navigation could also underlie non-spatial cognitive computations. Here, I review the anatomy, physiology and seemingly separate function of the hippocampal formation, and suggest how they can be combined into a coherent framework that gives rise to our ability to navigate in physical space, as well as in a more abstract knowledge space.

#### **1.4.1 Anatomy of the hippocampal formation**

The hippocampal formation is located in the medial temporal lobe (MTL) and consists of the hippocampus, the entorhinal cortex, the subiculum, the pre- and the parasubiculum. The three-layered hippocampus itself consists of the dentate gyrus and the Cornus ammonis (Latin: head of the ram), which has three subfields, CA1, CA2 and CA3. Inputs to the hippocampus are funnelled by the entorhinal cortex, which receives projections from most neocortical sites, in particular from perirhinal and parahippocampal cortices integrating information from association cortices (Lavenex and Amaral, 2000). Fibers from entorhinal layer III project to area CA1, and entorhinal stellate cells in layer II distribute their information via perforant path fibers to a large number of granule cells in the dentate gyrus. A granule cell only spikes if it receives simultaneous inputs from many entorhinal cells. Granule cells in the dentate gyrus, in turn, send so called 'mossy fibers' to CA3 pyramidal neurons. A typical CA3 neuron receives information from less than 50 granule cells.

However, because mossy fiber terminals are the largest axon terminals in the mammalian brain they are unusually powerful, and activity trains of a single granule cell can activate a post-synaptic CA3 pyramidal neuron under some circumstances (Henze et al., 2002). Granule cell projections distribute information widely across CA3, thereby orthogonalizing input patterns and performing pattern separation (O'Reilly and McClelland, 1994; Treves and Rolls, 1994). CA3 contains strong and largely random, excitatory recurrent connections from other CA3 neurons. In combination with a very high sensitivity to LTP, the divergent and convergent loops in CA3 make this substructure ideal for forming auto-associative memories, and for recovering previously stored patterns from partial cues. CA3 neurons also project 'Schaffer collaterals' to area CA1. Structures forming the hippocampal formation are therefore connected to each other via a unidirectional 'trisynaptic loop' (Figure 1.4). Most CA1 neurons target the subiculum and layer V of the entorhinal cortex, or the prefrontal cortex. While the subiculum projects largely to subcortical destinations, the entorhinal cortex targets neocortical areas, mostly perirhinal and parahippocampal cortices.



**Figure 1.4 Hippocampal anatomy and connectivity between subfields of the hippocampal formation.** Adapted with permission from Neves et al. (2008).

The hippocampal formation and MTL structures have dense reciprocal connections with the prefrontal cortex. CA1, the subiculum and the entorhinal cortex directly project to the prefrontal cortex (Condé et al., 1995; Jay and Witter, 1991; Swanson, 1981), and information is also relayed to vmPFC via the perirhinal cortex (Eden et al., 1992). Inputs from mPFC

and OFC reach the hippocampus mainly via projections to the entorhinal cortex and perirhinal or parahippocampal cortex, respectively (Lavenex and Amaral, 2000).

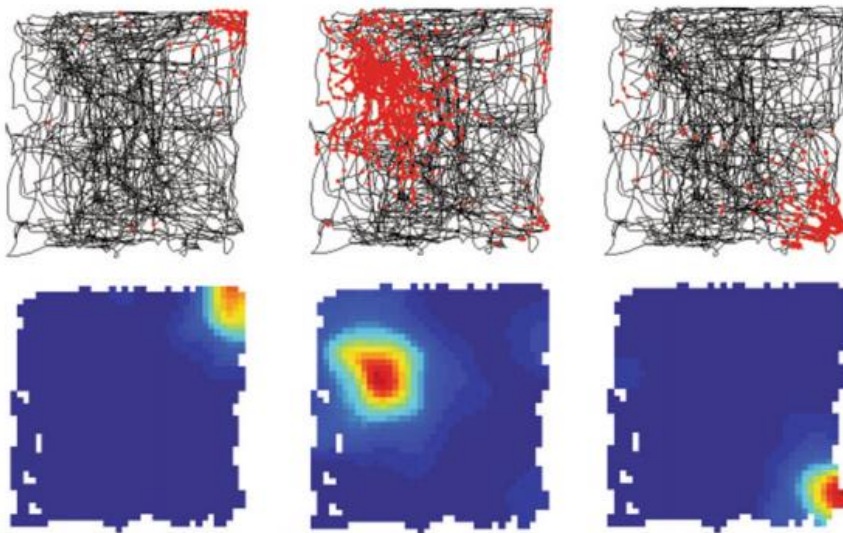
#### **1.4.2 Cognitive maps in the hippocampus and the entorhinal cortex**

Mammals are very successful at finding the shortest path to their home base after exploring an environment for food or searching for their missing pup, even in the absence of external sensory inputs (Mittelstaedt and Mittelstaedt, 1980). This type of ‘path integration’ requires the existence of a map that allows animals to keep track of directions and distances travelled during navigation.

In the early 1970s, it was discovered that such a map is in fact instantiated in the hippocampus. Neurons in hippocampal areas CA1 and CA3 were found to respond precisely to a circumscribed location in space (Figure 1.5) (O’Keefe, 1976; O’Keefe and Dostrovsky, 1971), such that their activity closely tracks the animal’s position during spatial navigation (Wills et al., 2010). More recently, place cells have also been recorded in human epilepsy patients (Ekstrom et al., 2003). ‘Place fields’, the location where such ‘place cells’ are maximally active, are distributed evenly across the entire environment. An animal’s location in space can thus be precisely decoded from the activity of a population of hippocampal place cells (Zhang et al., 1998). Firing rates of place fields are cone-shaped, and place cells respond equally independent of the direction from which an animal arrives in the place field (unless the animal navigates on a one-dimensional track). The distribution of place fields therefore provides an allocentric map of space that can be used for planning shortcuts or detours in spatial navigation (Leutgeb et al., 2005; O’Keefe and Nadel, 1978). No topographic relationship between the anatomical location of a place cell and its place field in physical space exists (Dombeck et al., 2010). In fact, if the animal moves to a new environment, the cognitive map completely reconfigures and a new, independent map is formed by rearranging the cells’ place fields (Muller and Kubie, 1987; Wilson and McNaughton, 1993). This remapping reverses if the animal returns to the original environment, providing the animal with stable and decorrelated representations of different environments. Thus, through hippocampal remapping, very large numbers of orthogonal cognitive maps can be created, allowing for the storage of vast amounts of independent memories. The orthogonalization is particularly strong in CA3, where neurons are densely connected via recurrent collaterals. If only small features of the environment are altered a

less radical rate remapping can be observed, where firing rates, but not place field locations are changed (Leutgeb et al., 2005).

Place cells are modulated by distant visual cues (Muller and Kubie, 1987), in particular geometric boundaries (O'Keefe and Burgess, 1996). Stretching or shrinking the dimensions of a testing apparatus leads to an elongation or shrinking of place fields in the same dimension, suggesting that place fields are arranged at fixed distances to boundaries or landmarks (O'Keefe and Burgess, 1996). The measure of allocentric distance to boundaries can be provided by border cells (Figure 1.7), firing at specific distances from geometric boundaries in the environment (Solstad et al., 2008).



**Figure 1.5 Place cell activity during spatial navigation.** Top: Black lines indicate the animal's path during foraging. Red dots indicate the location where three example CA3 place cells fired an action potential. Bottom: Firing rate map as a function of location. Adapted with permission from (Fyhn et al., 2007).

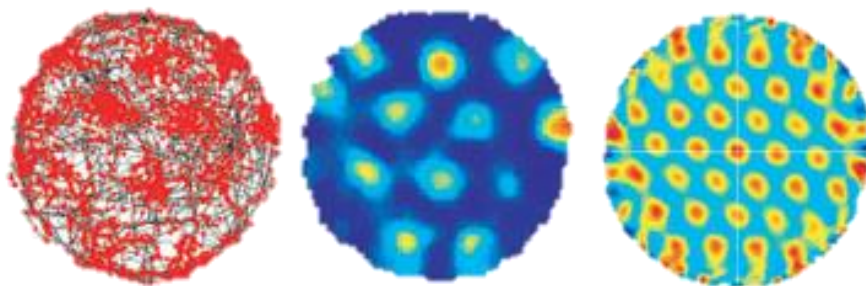
Place cells do persist in darkness if a dark period is preceded by a light period during which the animal can familiarize itself with an environment (Quirk et al., 1990). When an animal is immobilized and transported through space by the experimenter (Foster et al., 1989) or when an animal is stationary and the environment is rotated (Terrazas et al., 2005) place cells become largely inactive, suggesting self-motion cues are critical for their expression. Furthermore, place cell activity increases linearly with running speed even in situations where visual and vestibular inputs remain constant (Czurkó et al., 1999).

Additionally, experience and behavioural relevance influence place cell firing. For example, place cells differentiate between identical paths in space if the required action at the

end of the path differs (Wood et al., 2000), and the concentration of place cells is particularly high around reward locations (Hok et al., 2007; Hollup et al., 2001), a phenomenon which aids memory recall (Dupret et al., 2010). Furthermore, highly travelled routes are overrepresented in terms of the number of place cells encoding a location along this path. The organization of a cognitive map with respect to behavioural relevance is presumably mediated via the input the hippocampus receives from a wide range of other brain areas that are active during learning, including input from sensory cortices and dopamine projections from the midbrain. Hippocampal place cells therefore reflect a multimodal representation of the environment.

Other types of stereotyped cells are found in the medial entorhinal cortex, including border cells, head direction cells and grid cells. Entorhinal grid cells, the most abundant cell type in medial entorhinal cortex, are characterized by a very regularly spaced triangular firing fields (Figure 1.6, Hafting et al., 2005). The emerging hexagonal array spans the entire available environment and anatomically neighbouring grid cells typically show the same orientation and spatial frequency, but not the same phase. Furthermore, grid cell spacing decreases from ventral to dorsal entorhinal cortex, such that ventrally located grid cells have larger firing fields and larger spacing between adjacent firing fields than dorsally located grid cells, with discrete step-like increases in size (Stensola et al., 2012). This gradient mirrors the change in size of hippocampal place fields along the same axis, where place field sizes increase gradually from approximately 1 metre dorsally to up to 10 m ventrally (Kjelstrup et al., 2008). The dorsal-to-ventral axis in rodents corresponds to a posterior-to-anterior axis in humans (Insausti, 1993).

While the topography of place fields can take up to several days to form in a novel environment (Lever et al., 2002), grid patterns typically form as soon as an animal enters a novel environment and remain fixed thereafter. Crucially, grid cell patterns do not depend on visual input as they are also present in darkness (Hafting et al., 2005). Furthermore, grid cell firing is invariant to an animal's speed and behaviour (Hafting et al., 2005; McNaughton et al., 2006), suggesting it might provide a stable internal metric of the environment. Notably, a deformation of the environment causes complementary changes in grid scale (Barry et al., 2007) suggesting that grid patterns are modulated by visual input.



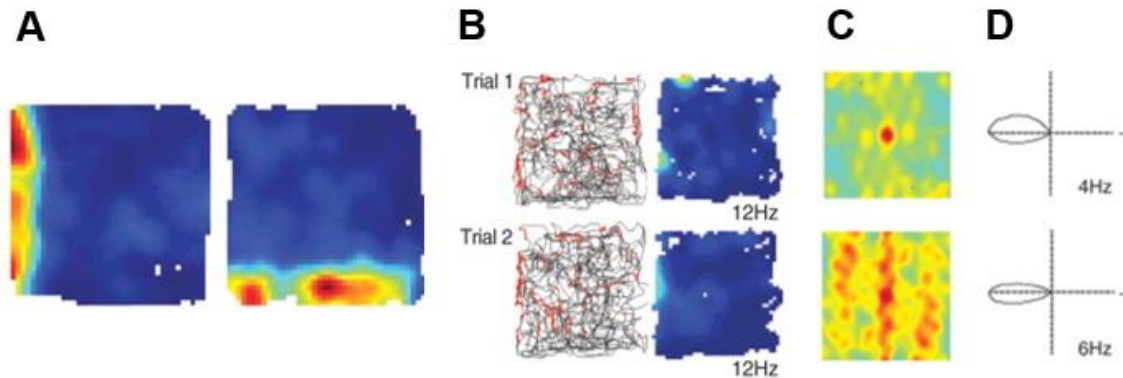
**Figure 1.6 Firing fields of a grid cell in the entorhinal cortex.** Left: the spiking activity of the grid cell (red dots) is overlaid on the animal's path in space (black line). Middle: Firing rate map. The peak rate for this neuron was 19 Hz. Right: Spatial autocorrelogram computed from the firing rate map. Adapted with permission from Hafting et al. (2005).

Grid cell patterns could result from self-motion cues indexing speed, and direction, and corrected by feedback from place cells and or boundary vector cells (McNaughton et al., 2006). Animals typically return straight to their starting location after leaving their nest to find food, even in the absence of sensory cues. This is taken as evidence for a path integration mechanism, whereby self-motion cues are continuously tracked such that the shortest path can be computed to return to the starting point. Path integration requires a navigation system assessing speed, elapsed time, head direction and initial position (Buzsáki and Moser, 2013). Boundary cells (Solstad et al., 2008) as well as visual information entering MEC from parahippocampal and postrhinal cortex (Epstein et al., 2007) ensure alignment of the path integration signal with the external world.

Place cells (Ekstrom et al., 2003) and grid cells (Jacobs et al., 2013) have also been recorded in humans epilepsy patients using electrophysiological recordings. Furthermore, because the phases of grid cells are aligned, a six-fold rotational symmetry can be detected in the fMRI BOLD signal in the entorhinal cortex when human subjects navigate through a virtual environment (Doeller et al., 2010). This technique also provided evidence for grid cells in mPFC, parietal cortex and temporal cortex. Furthermore, recent evidence shows that the entorhinal cortex encodes Euclidian distances to goals (Howard et al., 2014) and goal direction (Chadwick et al., 2015).

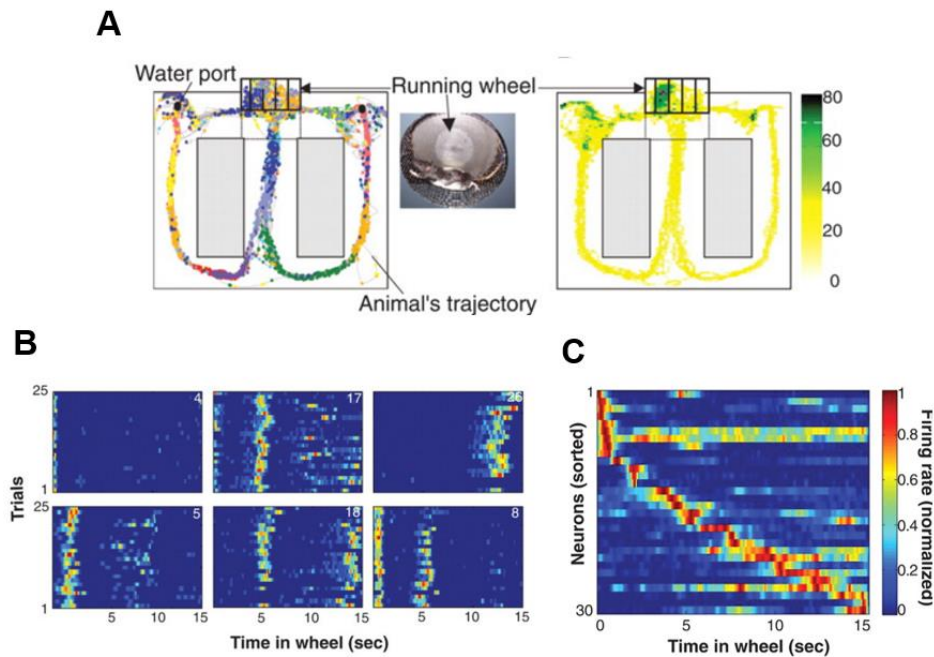
Some entorhinal grid cells are directly modulated by head direction (Sargolini et al., 2006), providing the hippocampal formation with directional information. Head direction cells, firing specifically when an animal's head faces a certain direction in the environment irrespective of the animal's location, were first identified in the presubiculum by Jeffrey Taube and James Ranck (Figure 1.7B-D, Taube et al., 1990). Since the initial discovery, 'head

direction cells' have been identified across a wide range of brain regions, including in the thalamus (Mizumori and Williams, 1993) and the entorhinal cortex (Sargolini et al., 2006).



**Figure 1.7 Examples of medial entorhinal border cells and head direction cells.** **A** Colour-coded firing rate maps for two representative example border cells, recorded in medial entorhinal cortex. Adapted with permission from Solstad et al. (2008). **B** Left: the spiking activity of the cell (red dots) is overlaid on the animal's path in space. Right: Colour-coded firing rate map. The peak rate for both neurons was 12 Hz. **C** Colour-coded spatial autocorrelogram of two head-direction cells, computed from the firing rate map. **D** Polar plots visualizing directional tuning of the firing rate. Firing rate as a function of head direction. Adapted with permission from Sargolini et al. (2006).

Another dimension encoded in the hippocampal formation is the temporal order of events. Only recently hippocampal 'time cells' have been discovered, which fire at specific moments during temporally structured experiences irrespective of an animal's spatial location. The neural ensemble as a whole thereby signals the passage of time. For example, when a rat runs on a running mill where it maintains a stable position in space, time cells responded at specific segments of the run (Figure 1.8) (Pastalkova et al., 2008). Time cells also keep track of the time that elapses between events in a task where rats learn to associate objects with odours presented after a delay (MacDonald et al., 2011). Critically, in this setting the ensemble activity is stable over subsequent experiences, but differs depending on which information is required for memory recall. Similarly, recordings in primate hippocampus have revealed the existence of time cells in a temporal-order memory task (Naya and Suzuki, 2011). Time cells could provide the temporal context of episodic memories and add to the notion that the hippocampus links event sequences in space and time.



**Figure 1.8 Hippocampal time cells.** **A** Colour-coded spikes of multiple hippocampal CA1 neurons. Dots indicate the location where place cells fired an action potential. In between runs around the maze the animal was required to run in the wheel. **B** Colour-coded firing rate of six neurons, recorded during wheel running. Each line corresponds to a separate trial. Time cells fire at specific delays during the wheel running episode. **C** Colour-coded firing rate for a set of neurons recorded during wheel running, sorted by latency of peak firing rate. The neuronal ensemble encodes the time spent on the wheel. Adapted with permission from Pastalkova et al. (2008).

### 1.4.3 Hippocampal place cells and entorhinal grid cells

The combined information concerning speed, elapsed time, direction, position and boundaries provided by cells in the hippocampus and entorhinal cortex can provide a distance metric and directional reference frame for mapping the environment. This map allows an animal to establish its exact position in space and plan trajectories through the environment. However, the exact relationship between hippocampal place cell and entorhinal grid cell patterns is still a matter of debate. Grid cells can be found in superficial and deep layers of the entorhinal cortex, suggesting that they process inputs to, and outputs from, the hippocampus. Most entorhinal cells in layer II and III, which project to the hippocampus, have grid cell properties (Sargolini et al., 2006). Furthermore, MEC contains head direction and border cells, and it is ideally placed to integrate spatial information with inputs it receives from neocortical areas (Hafting et al., 2005). It has therefore traditionally been assumed that MEC neurons project information about spatial location, direction and distance to place cells, with hippocampal place fields constituting a ‘read-out’ of entorhinal grid cells. Indeed, a linear combination of multiple grid cells with various spatial frequencies



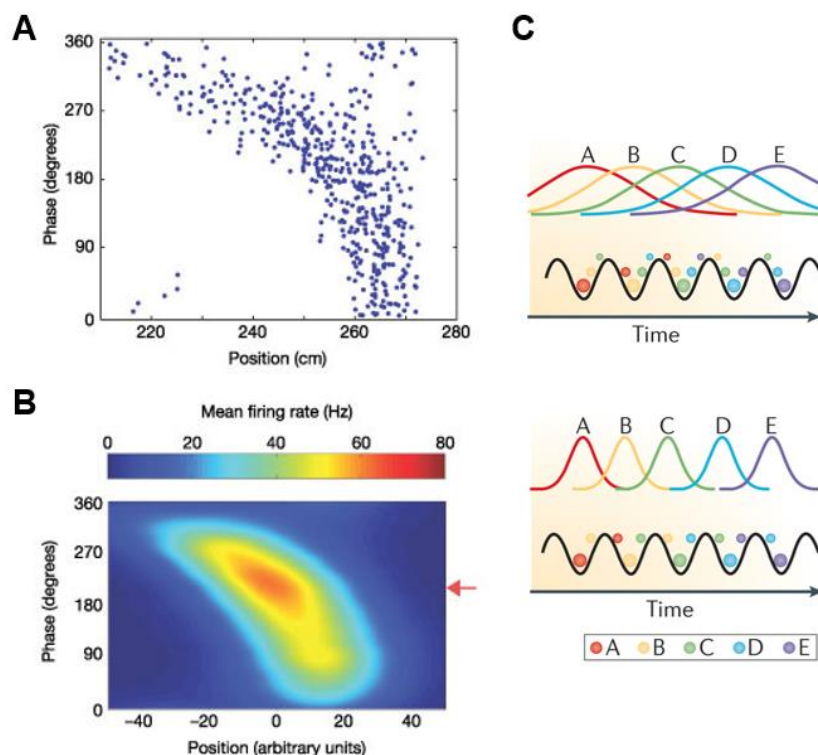
and random phases could provide a precise estimate of an animal's location, suggesting that grid cell activity could constitute a basis set that can be combined linearly to generate place fields in the hippocampus (Fuhs and Touretzky, 2006; McNaughton et al., 2006). A place field in this setting emerges at the location where most grid cells are in phase. Theoretically, it has been demonstrated that the summation of grid cells with biologically plausible differences in grid frequency, aligned phases, and random grid orientation would indeed lead to the emergence of place cell-like activity (Solstad et al., 2006). Place cells in this context would not need to be hard-wired, but could instead emerge from Hebbian plasticity between random grid cell-place cell connections if grid cells vary sufficiently in orientation, phase and scale (Rolls et al., 2006). This is also in line with the observation that place fields in CA1 remain unaffected when the input from hippocampal CA3 to hippocampal CA1 is removed (Brun et al., 2002), suggesting either that the perforant-path input from the entorhinal cortex provides CA1 with spatial information, or this information is directly computed within CA1.

However, recent evidence is inconsistent with the notion that place cells are downstream read-outs of grid cell activity. Size and shape of place cell firing fields are mostly unaffected by the absence of grid cell inputs (Hales et al., 2014; Koenig et al., 2011) whereas inactivating the hippocampus leads to a loss of hexagonal grid cell firing (Bonnievie et al., 2013). In this framework, the hippocampus can be considered the area encoding individual experiences or locations in space, whereas the entorhinal cortex computes the relationship between experiences, locations, objects or other types of relational information. In other words, grid cell activity could represent a read-out of place cell activity, whereby spatial and contextual information arising from external visual inputs and computed by the hippocampus can be processed and efficiently signalled to relevant cortical areas (Barry et al., 2006). Indeed, during development, place cells mature before grid cells (Langston et al., 2010; Wills et al., 2010) and the development of grid cells coincides with an increased accuracy of place cell activity (Muessig et al., 2015).

#### **1.4.4 Sequence coding by theta phase precession**

Population activity in the hippocampus is characterised by large oscillatory activity in the theta frequency band (7-12 Hz), which is present during ongoing behaviour such as spatial navigation (Green and Arduini, 1954). The intracellular membrane potential of individual place cells, however, oscillates faster than this population rhythm. As a consequence, their spikes shift relative to the phase of the theta oscillations as an animal transverses a place field,

a phenomenon called ‘phase precession’ (Figure 1.9A,B). When a rat first enters the firing field of a place cell, the place cell’s activity occurs at a late theta phase, and systematically shifts to earlier theta phases with the progression through the firing field (O’Keefe and Recce, 1993). Phase precession can also be observed in entorhinal grid cells, even if hippocampal inputs are removed (Hafting et al., 2008), and in medial prefrontal cortex (Jones and Wilson, 2005). The animal’s location in space and the phase of neuronal spiking are therefore correlated, providing an animal with a precise temporal code of its position.



**Figure 1.9 Hippocampal phase precession.** **A** Neural activity as a function of position and phase. Dots represent location and theta phase at the time when an action potential is fired in a CA1 place cell. **B** Firing rate of the neuron in **A** as a function of position and phase. **A** and **B** reproduced with permission from Mehta et al. (2002) **C** Schematic illustrating the compression of sequences through phase precession in ventral/anterior hippocampus (top) and dorsal/posterior hippocampus (bottom). Coloured Gaussians represent hippocampal place fields. Coloured circles represent spikes of the corresponding place cell. As the animal transverses a cell’s firing field, the spiking activity precesses. As a consequence, within a theta cycle, the sequence (A-E) is preserved and compressed. Importantly, more ventral parts of the hippocampus can accommodate longer sequences. Reproduced with permission from Strange et al. (2014).

Furthermore, the spikes of neurons whose place cells overlap occur within one theta cycle in an order that preserves the temporal relationship between place fields encountered during navigation (Figure 1.9C). Not only does the temporal relationship of spikes provide the animal with an estimate of the distance travelled at a millisecond time scale (Skaggs et al.,

1996), phase precession also compresses the place cell activity enough to allow spike-timing dependent plasticity to occur between the corresponding place cells. This plasticity between sequentially activated place cells might underlie the formation of cell assembly sequences or episodes (Blum and Abbott, 1996). When animals first encounter a novel environment, they explore in random directions, and the paths they take often cross (Whishaw and Brooks, 1999). The locations where paths cross are thus part of multiple independent routes, which become linked through Hebbian plasticity, ultimately resulting in the formation of a map. At the same time, the temporal compression results in a concurrent representation of present, past and future locations. The current location is represented by the neurons that are active at the trough of the theta oscillation. However, at the same time assemblies active at descending and ascending phases correspond to passed and upcoming locations. As a consequence, the current location is always embedded in a spatiotemporal sequence.

#### **1.4.5 Map-based influences on behaviour**

During slow wave sleep (SWS) and rest periods after performing a task, self-organized sharp wave ‘ripple’ oscillations emerge in CA3 place cells, and drive CA1 pyramidal cells (Csicsvari et al., 2000). The sequence in which place cells are activated during sharp-wave ripples is not random. Instead, firing sequences of hippocampal place cells correspond to recent spatial experience, which is replayed at an accelerated speed (Diba and Buzsáki, 2007; Karlsson and Frank, 2009; Skaggs and McNaughton, 1996; Wilson and McNaughton, 1994). This selective and repeated activation of cell assemblies could underlie memory consolidation by aiding synaptic plasticity in the hippocampus as well as in neocortical targets. Disrupting sharp wave ripples during sleep leads to memory impairments on a spatial task, emphasizing the causal relationship between ripples and memory consolidation (Girardeau et al., 2009).

Rapid reactivation of spatial sequential activity patterns can also be observed in non-exploratory wake periods immediately after spatial experience, in particular in novel environments (Foster and Wilson, 2006). Critically, these replay events typically occur in reverse temporal order and they can occur at the same time as a VTA prediction error signal if a reward has been experienced (Gomperts et al., 2015). Reverse replay in combination with a VTA prediction error signal can solve the temporal credit assignment problem, by facilitating the propagation of value information from a rewarded location backwards along an experienced trajectory. This is also consistent with the observation that place preferences

can be experimentally induced by pairing a replayed trajectory with a rewarding stimulation of the medial forebrain bundle (de Lavilléon et al., 2015).

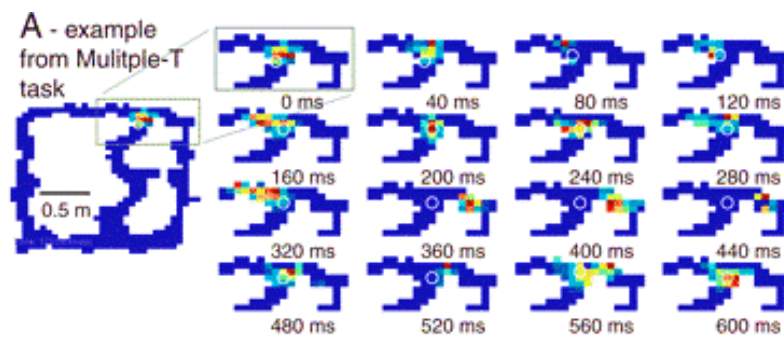
Replay can also contribute to integrating multiple experiences into a coherent representation. Due to the fast time scale of sharp-wave ripples, representations that occur at longer intervals can be compressed into a suitable time scale for synaptic plasticity. Indeed, after exploring a large track, multiple sharp wave ripple events are chained such that large trajectories are reactivated in a coordinated fashion, including trajectories distant from an animal's current location (Davidson et al., 2009). In this way, replay putatively contributes to the creation of a unified representation of an extended experience. This could also allow unnecessary details of single experiences to be discarded, and regularities and structure across experiences extracted across experiences. It also allows for new information to be combined with previously acquired and reactivated knowledge.

It is assumed that memories are encoded in a distributed fashion across associative areas of the neocortex that are active during learning (Cowansage et al., 2014). These populations of neurons form a 'memory engram', i.e. a cellular representation of a memory that, if reactivated, reinstates the corresponding memory experience. The role of the hippocampus is to reinstate these distributed memories during memory recall (Tanaka et al., 2014) and integrate new information during learning. It is therefore not surprising that the sharp-wave-ripple consolidation processes involve interactions between a range of brain areas, including the hippocampal-entorhinal cortex, prefrontal brain areas as well as association cortices (O'Neill et al., 2008). For example, simultaneous recordings in the hippocampus and the primary visual cortex during sleep suggest that sequence reactivation occurs simultaneously in both in hippocampus and visual cortex (Ji and Wilson, 2007).

Replay phenomena also have a corresponding counterpart in *preplay* events, where hippocampal activity sweeps through state spaces that correspond to potential future trajectories. Preplay can also be observed during sharp wave ripple oscillations, where sequences are often biased to progress from the animal's current location to a goal location. Crucially, the decoded sequences are predictive of an animal's future trajectory (Pfeiffer and Foster, 2013). This phenomenon may correspond to the neural instantiation of prospective search, demonstrating that a cognitive map may indeed support flexible model-based planning in goal-directed behaviour. The delivery of a reward in a visible, yet unexplored environment can also result in preplay of the corresponding trajectory in space, demonstrating goal-directed future simulation in rats (Ólafsdóttir et al., 2015). Prospective

planning activity involving the reactivation of reward states can also be seen when human participants perform model-based choices in the two-step task (Doll et al., 2015).

Preplay has also been reported in the theta-frequency domain. When rats face a choice to turn left or right at a decision point in a T-maze, hippocampal activity no longer reflects the animal's current location in space. Instead, the sequence in which hippocampal place cells have been activated when the animal previously experienced the left or the right path is preplayed in the theta frequency domain (Johnson and Redish, 2007). It thus appears as though the animal is contemplating the choice options at a decision point by projecting themselves into the future. It is worth noting, however, that no direct link between preplay events in theta and subsequent choice was observed in this study, indicating that the preplay signature may instead reflect upcoming states rather than a signature of goal-directed planning.



**Figure 1.10 Forward-projecting neural representation at a choice point.** When the animal reaches a decision point, neural activity in the hippocampus no longer reflects the animals' actual location in space, but the representation instead moved ahead into the two arms of the T-maze. Reproduced with permission from Johnson and Redish (2007).

#### 1.4.6 Cognitive maps in non-physical abstract space

As humans, we live in ever-changing world and numerous random events occur at all times. While most of these events are insignificant for our survival, storing some information can be important for processing future sensory experiences and modifying our behavioural policies. The relationships between events, objects and other types of information is particularly relevant for goal-directed planning. Just like relationships in physical space, these relationships can often be defined in terms of distances and positions, e.g. semantic proximity (Trope and Liberman, 2010), social proximity, or even proximity to myself in the past or the future. Similarly, paths in physical space can be considered analogous to episodes of sequentially experienced events or the repeated experience of semantic relationships. Storing

abstract relationships as a cognitive map would have tremendous advantages for goal-directed behaviour, as a network of memories or conceptual relationships would allow us to travel new routes through an abstract memory space and solve novel problems (Eichenbaum and Cohen, 2014).

There is now good evidence that the brain networks underlying spatial navigation (Doeller et al., 2010), memory (Binder et al., 2009), imagination (Schacter et al., 2012) or valuation (Clithero and Rangel, 2014) overlap substantially. All of these processes require binding of disparate details into coherent events, which can be recollected as a whole when a cue is encountered (Buzsáki and Moser, 2013). As outlined above, the hippocampal formation is ideally suited for storing vast amounts of experiences and environments, constructing relational information and organizing experiences into event sequences. It therefore provides an ideal basis for representing non-spatial relational knowledge (Buzsáki and Moser, 2013).

In recent years, evidence has accumulated that the hippocampus computes non-spatial and abstract relational mappings between arbitrary stimuli. In rats, the response of individual neurons in rat CA1 differs depending on whether an odour is presented at a position consistent with its position in a previously learned sequence or not (Allen et al., 2016), suggesting that the hippocampus also encodes the temporal relationship between non-spatial stimuli. This is corroborated by a recent neuroimaging study in humans demonstrating that the hippocampus encodes the temporal order of objects in learned object sequences (Hsieh et al., 2014).

Furthermore, rats can infer a link between odours A and C after learning that odours A and B as well as odours B and C are associated (Bunsey and Eichenbaum, 1996). This suggests the rats construct cognitive maps of simple associations allowing them to make transitive inferences. Critically, animals with hippocampal lesions are specifically impaired in forming this transitive link. In humans, a signature of transitive inference can be found in anterior, but not posterior, hippocampus (Collin et al., 2015; Horner et al., 2015; Preston et al., 2004), suggesting that this part of the hippocampus generalizes over individual episodes to facilitate inferential reasoning (Komorowski et al., 2013). This is in line with the observation that the size of place fields increases from posterior to anterior in the human hippocampus, or along the dorso-ventral axis in rodents. As a consequence, neurons with more distant place fields in the anterior (or ventral) hippocampus are more likely to fire together, resulting in the formation of higher-order links between distant locations or stimuli, e.g. between locations A-E in the schematic example (Figure 1.9C). Such higher-order links

between place cells firing at non-adjacent places could putatively underlie transitive inference (Strange et al., 2014).

More generally, the human hippocampus encodes the statistical relationships between non-spatial stimuli in the environment (Schapiro et al., 2012), and clusters mutually predictive stimuli into event representations (Schapiro et al., 2013). There is also some evidence that non-spatial relational information is organized in a map in the hippocampus, as demonstrated in an fMRI experiment investigating the representation of social relationships (Tavares et al., 2015).

## 1.5 Thesis overview

In this thesis, I set out to investigate basic mechanistic principles underlying learning and decision making by combining computational with representational techniques in fMRI. In a first set of studies, I measured changes in similarity between neuronal representations during learning using repetition suppression. Because different neuronal computations are performed by overlapping neural circuitries, learning to perform one behaviour can cause plasticity in an overlapping computation. When subjects learn the preferences of a new individual, repetition suppression is observed between the subjects' representation of themselves and their representation of this newly learnt partner. Strikingly, this new representational overlap predicts a change in subjects' own preferences. In fact we show that subject whose preferences are most influenced by others are those who also develop the most overlapping neural representations. This suggests that social influences can act at the most basic level of neuronal representations.

In a second set of studies, I use a similar representational approach in combination with an implicit learning paradigm to investigate how the brain extracts information about statistical regularities and organizes this information to build models of the world. Goal-directed behaviour requires a neural representation of the associations between objects, events and other types of information. The mechanisms underlying the association of pairs of objects are well characterized, and involve an increase in the similarity of the respective object representations. Much less is known about how the human brain stores multiple associations that form a more complex global structure. Does the cortex continue to store simple associative links, or is global knowledge about the relationship between objects that have not been directly associated nevertheless incorporated in the representation? Here, I

address this question from a functional perspective and ask whether a signature of a global structure is apparent following an implicit learning paradigm. I find that humans acquire implicit knowledge about global structure and store this knowledge in an entorhinal map. These effects can be explained by two mechanistic models: Firstly, a map emerges in a simple Hopfield network with auto-associative attractors and Hebbian plasticity between associated objects. This suggests that global knowledge about the relationship between non-associated objects emerges through increases in representational similarity for pairwise associations. I also propose a conceptually distinct model, wherein the entorhinal cortex performs an eigenvalue decomposition of place cell activity. This model not only explains the emergence of a global structure in the entorhinal cortex, and it can also account for the characteristic hexagonal arrangement of grid cell firing fields.



## 2 METHODS FOR INVESTIGATING PHYSIOLOGICAL BRAIN ACTIVITY AND BEHAVIOUR

\* The material presented in Chapter 2.4 is partly taken from a paper being prepared for publication as the following article:

Barron, HC<sup>+</sup>, Garvert, MM<sup>+</sup> & Behrens TEJ (in press). Repetition suppression: a means to index neural representations using BOLD', *Philosophical Transactions of the Royal Society B*

<sup>+</sup> equal contribution

The ideas were developed jointly with Helen Barron, and it is not possible to clearly identify individual contributions.

## 2.1 Introduction

A popular technique used to image physiological activity in the human or animal brain with a high spatial resolution is blood oxygenation level dependent (BOLD) fMRI (Ogawa et al., 1990). In fMRI, brain activity is measured indirectly as a change in blood flow thought to accompany neural activity. Oxygenated and deoxygenated blood differ in their magnetic susceptibility, which can be measured non-invasively using the phenomenon of nuclear magnetic resonance (NMR, Bloch, 1946; Purcell et al., 1946).

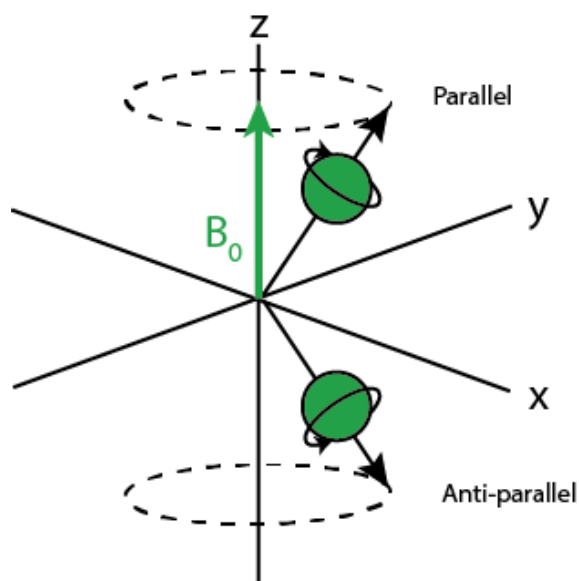
Typically, fMRI data has been analysed in a mass univariate manner, which permits making inferences about brain activity at a macro-anatomical scale, i.e. at the level of brain regions. This has led to important contributions to our understanding of regional specialization in the brain, such as the identification of the parahippocampal place area (Epstein and Kanwisher, 1998) or the fusiform face area (Kanwisher et al., 1997). More recently, techniques such as fMRI adaptation have been developed that can be used to measure activity at a meso-scale and provide a more fine-grained access to neural representations and coding schemes. In this chapter, I will first review the general principle of fMRI, and subsequently discuss mechanisms and applications of fMRI adaptation.

## 2.2 Principles of Magnetic Resonance Imaging (MRI)

### 2.2.1 Static magnetic field and magnetization

An MRI scanner generates a strong static magnetic field of typically 1.5 to 7 Tesla (T), denoted  $B_0$ . This magnetic field induces a proton spin flip in atomic nuclei that have a magnetic moment such as the hydrogen nuclei found in water molecules in the human brain. Outside a magnetic field the nuclei are randomly oriented, but in a magnetic field they align parallel or anti-parallel to the direction of the magnetic field, also denoted z-axis (Figure 2.1, Figure 2.2A). The degree of alignment depends on the strength of the magnetic field. Since slightly more protons align parallel, this results in a net magnetization ( $M$ ) along the longitudinal axis  $B_0$ , but no magnetization along the transverse plane. Protons precess around the z-axis at a random phase, but a specific angular frequency, called ‘Lamor frequency’. This frequency is determined by the field strength  $B_0$  and by an atom-specific constant  $\gamma$  called the gyro-magnetic ratio (42.58 MHz/T for hydrogen):

$$\text{Larmor frequency: } f = \gamma B_0 \quad (2.1)$$



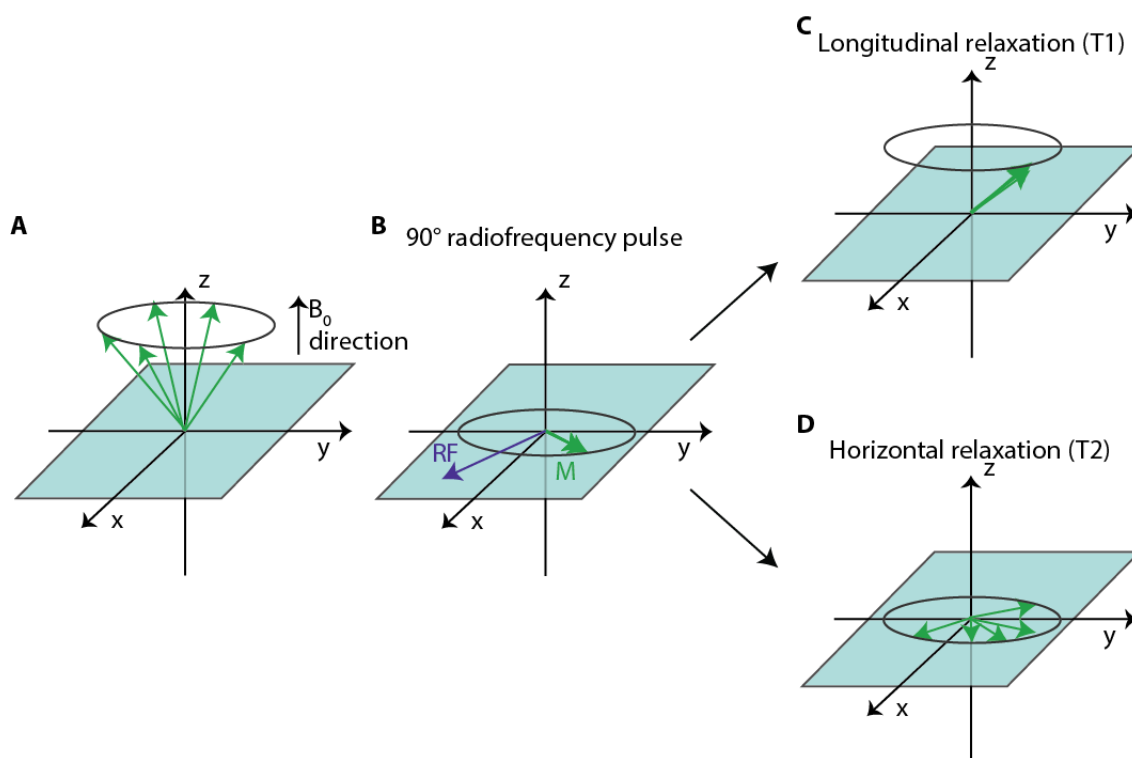
**Figure 2.1** Hydrogen atoms align with the static magnetic field and precess about the z-axis.

### 2.2.2 Application of a radiofrequency pulse

Electrical coils placed around the participant's head deliver oscillating radiofrequency (RF) pulses, also called  $B_1$  field, which is perpendicular to the static magnetic field  $B_0$  (Figure 2.2B). When energy is delivered at the resonance frequency as determined by the Larmor equation, the targeted molecules (here: hydrogen molecules) absorb energy and jump from a low-energy state to a high-energy state. As a consequence, the net magnetization vector  $M$  tips towards the transverse x-y plane (Figure 2.2A-B) and a new transversal magnetization is established, while the longitudinal magnetization decreases. Oscillations in the x-y plane generate an electromagnetic signal which can be measured by the receiver coils. Note the net magnetization need not be flipped by  $90^\circ$ , the flip angle (final angle between  $B_0$  and  $B_1$ ) can be smaller.

When the radiofrequency pulse is removed, the atomic nuclei lose energy and exponentially decay back to the direction of  $B_0$ . The longitudinal relaxation corresponds to the restoration of the net magnetization along  $B_0$ , and its time course is characterized by the time constant  $T_1$  (Figure 2.2C). At the same time, the transverse magnetization decays exponentially (transverse relaxation), and the atomic nuclei dephase in the x-y plane at a time course characterized by the time constant  $T_2$  (Figure 2.2D). The decay in transverse magnetization is caused by molecular interactions between neighbouring nuclei (spin-spin



interactions). Through these interactions, nuclei pass energy from one to another such that their rotations become desynchronized. A second important component affecting T2 relaxation are local magnetic field inhomogeneities. The parameter T2\* (“apparent T2”) constitutes the time constant capturing both effects, and is the basis of fMRI. Crucially, T1 and T2 depend on the proton density in a given tissue. T1 for white matter is much shorter than T1 for grey matter or cerebral spinal fluid (CSF), and also T2 differs between CSF and grey / white matter. As a consequence, the time constants can be used to generate images of different contrasts.



**Figure 2.2** The consequences of applying a radiofrequency pulse to a  $B_0$  field. **A** Before applying a pulse, the atomic nuclei precess about  $B_0$  at a frequency determined by the Larmor equation. **B** A radiofrequency (RF) pulse perpendicular to  $B_0$  tips the net magnetization vector  $M$  towards the transverse x-y plane. **C** When the radiofrequency pulse is removed, the longitudinal magnetization decays back to  $B_0$  (T1 relaxation). **D** Furthermore, the atomic nuclei dephase in the x-y plane (T2 relaxation).

The contrast of the image obtained with MRI is influenced by the echo time (TE), the repetition time (TR, Table 2.1), as well as the flip angle and the strength of the RF pulse. TE denotes the time between the application of a RF pulse, which causes the flip of the magnetization vector into the x-y plane, and the measurement of the electromagnetic signal. The longer the TE, the more T2 relaxation. TR corresponds to the time between two successive RF pulses. The longer the TR, the more magnetization is available for the next

excitation. Very long TRs measure the proton density of a voxel, but are suboptimal if the contrast between tissue types is of interest, as all tissue types ultimately recover maximal longitudinal magnetization. T1-weighted sequences used for anatomical imaging use short TEs and short TRs ( $TE < 30\text{ms}$  and  $TR < 1\text{s}$ ). Longitudinal magnetization recovers faster for white matter than for grey matter, such that white matter appears brighter on T1 weighted images (Table 2.1).

		<b>TE</b>	
		<b>Short</b>	<b>Long</b>
<b>TR</b>	<b>Short</b>	<b>T1</b> 	<b>Not used</b>
	<b>Long</b>	Proton density (not used in this thesis)	<b>T2</b> 

**Table 2.1 Image contrast depending on repetition time (TR) and echo time (TE).**

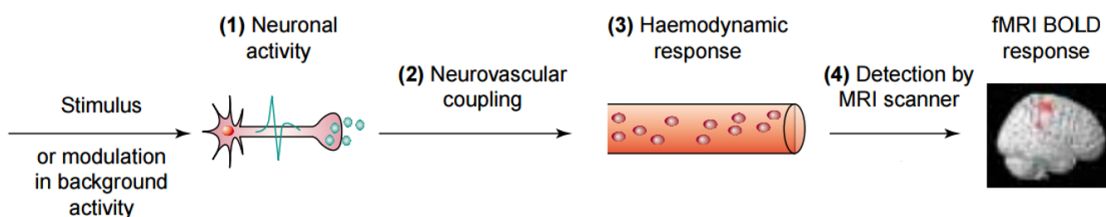
### 2.2.3 Field gradient

To locate atoms in a sample, a gradient amplifier applies an additional z field, or “field gradient”, that varies linearly along the x-axis. As a consequence, the atomic nuclei along the x-axis precess at different frequencies, because the precession frequency depends linearly on the magnetic field strength (see Larmor equation, (2.1). The signal picked up by the receiver coil is then a linear combination of individual frequencies, and spatial information about the obtained signal can be obtained by matching the RF pulse to the Larmor frequency of a particular location. The entire brain can be sampled by applying RF pulses of varying frequency. The measurements are acquired in the frequency domain (k-space).

## 2.3 Functional magnetic resonance imaging (fMRI)

### 2.3.1 The BOLD signal

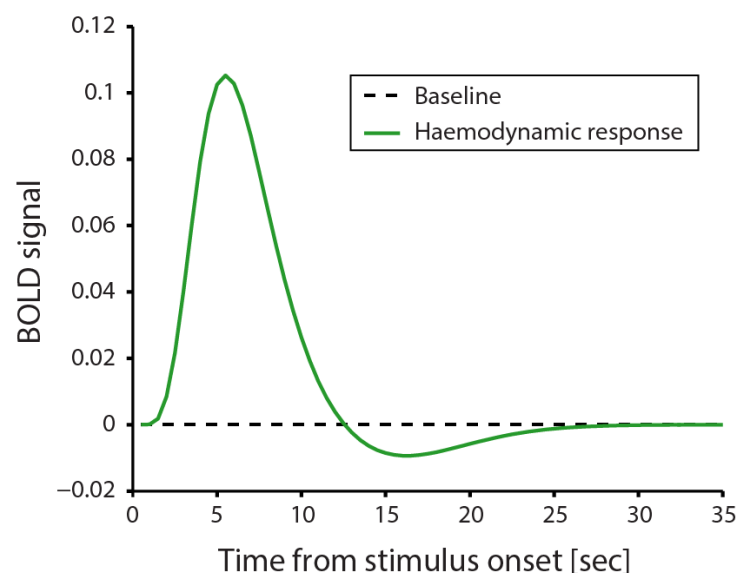
fMRI measures neuronal activity indirectly through its relationship with regional changes in blood flow (Figure 2.3). Following neuronal activity, ionic gradients need to be restored and neurotransmitters need to be recycled. These processes demand energy provided in the form of adenosine triphosphate (ATP). While a small amount of ATP can be produced by anaerobic glycolysis, under healthy conditions 90% of ATP is produced by oxidative phosphorylation. This process is aerobic, and requires oxygen as well as glucose. As the brain has no available pool of either, both are delivered via the cerebral vascular system. Since energy requirements are particularly high during neural activity, cerebral blood flow and vessel dilation increase locally under high neuronal activity. However, cerebral blood flow surpasses oxygen consumption such that areas with an increase in neural activity display a net increase in the amount of oxygen present.



**Figure 2.3 Determinants of the BOLD signal.** Neuronal activity leads to a regional increase in the demand for oxygen. Complex neurovascular coupling processes then trigger a haemodynamic response, namely an increase in local blood flow and vessel dilatation. This changes the relationship between the amount of oxygenated and de-oxygenated blood present in a region, which can be detected by the MRI scanner and ultimately results in the fMRI BOLD response. Adapted with permission from Arthurs and Boniface (2002).

Oxygen is transported via haemoglobin, an iron-containing molecule in erythrocytes (red blood cells). When oxygen is released, oxygenated haemoglobin becomes deoxygenated. Crucially, oxygenated and deoxygenated haemoglobin differ in their magnetic susceptibility. Whereas oxygenated haemoglobin is diamagnetic, deoxygenated haemoglobin is paramagnetic because it contains two unbound iron-containing haem groups (Pauling and Coryell, 1936). These haem groups alter the magnetic field by introducing local field inhomogeneities, thereby causing dephasing of spins of nearby protons. This accelerates the decay of the transverse magnetization and thus shortens the  $T_2^*$  time. As a consequence, oxygenated and deoxygenated produce different  $T_2^*$  signals, which can be contrasted.

The dynamics of neurovascular coupling are slow and changes in blood flow substantially lag behind neuronal activity. Initially, oxygen demand results in a small increase of deoxygenated haemoglobin, resulting in a delayed onset or even a small initial dip of the stimulus-induced BOLD response (Figure 2.4). Then, vasodilatation takes effect resulting in an oversupply of oxygen and a decrease in deoxygenated haemoglobin. This phase typically reaches its peak only after 6-10 sec (Logothetis, 2003). After about 20 sec, a peak in the BOLD response is followed by an undershoot. These slow dynamics of the hemodynamic response impose important constraints on the temporal resolution of fMRI. Whereas neuronal activity typically occurs on the millisecond timescale, the slow hemodynamic response function (HRF) acts as a temporal filter for neural activity. Fast neuronal activity can therefore not be detected in the BOLD response. It is also important to note that the exact dynamics of the BOLD response differ between brain regions and species.



**Figure 2.4** The hemodynamic response function.

### 2.3.2 Neurophysiological basis of the BOLD signal

Simultaneous recordings of the BOLD signal, neuronal spiking activity as well as LFPs in the macaque visual cortex have demonstrated the BOLD response correlates better with LFPs than with a region's spiking activity (Logothetis et al., 2001). This is consistent with the observation that manipulations of neuronal firing activity do not cause changes in cerebral blood flow (Lauritzen and Gold, 2003; Thomsen et al., 2004). LFPs reflect the summated excitatory and inhibitory postsynaptic potentials, i.e. mostly incoming synaptic activity and the result of local cortical computations rather than outgoing spiking information, although

LFPs and spiking activity usually correlate (Logothetis, 2003; Logothetis et al., 2001). This should be kept in mind when comparing fMRI studies to neurophysiological recordings in animals which mostly measure firing activity.

In typical fMRI experiments, the hemodynamic response increases linearly with neural activity (Li and Freeman, 2007; Logothetis, 2003; Rees et al., 2000). However, nonlinear threshold and saturation effects have also been described (Sheth et al., 2004). Such nonlinear effects can, for example, be observed if stimuli are spaced close together in time and are thought to be related to neuronal refractoriness or hemodynamic saturation effects (Friston et al., 1999). The practical implication is that this mandates the use of inter-trial intervals of a few seconds to ensure that the assumption of linearity holds.

### **2.3.3 fMRI data analysis**

In my experiments, I collected functional brain data using a T2\*-weighted echo-planar imaging (EPI) sequence, a T1-weighted whole-brain structural MRI scan and individual fieldmaps for each subject. To allow for scanner equilibration, the first five volumes of each experimental block were discarded as “dummy volumes”. Subsequently, both the fMRI and the structural scans were spatially preprocessed before performing statistical analyses on the data. This reduces the inter-subject variability in brain anatomy by spatially transforming each subject’s images to a standard anatomical space and therefore increases validity and sensitivity in group analyses (Friston et al., 1995). Furthermore, during the preprocessing procedure various sources of noise, which corrupt the fMRI signal, were corrected. These include movement-induced noise during scanning, physiological parameters (heartbeat and respiration) and low frequency signal drift. After applying preprocessing procedures, parametric statistical models can be designed to test hypotheses about the relationship between experimental parameters and BOLD activity.

All procedures reported in this thesis were performed using Statistical Parametric Mapping software (Wellcome Trust Centre for Neuroimaging, UCL, <http://www.fil.ion.ucl.ac.uk/spm/>) implemented in Matlab R2012b (The Mathworks, USA).

### **2.3.4 Slice time correction**

Since EPI volumes are collected in ascending order, the first slice in a volume can be collected up to 1 TR later than the last slice in a volume. To correct this, a reference slice is



chosen (typically a middle slice in the sequence) and a phase shift is introduced which shifts the Fourier transform of each voxel's time series such that they have the values they would have had they been sampled at the same time as the reference slice.

### **2.3.5 Bias-correction**

The 32-channel head coil can cause local field inhomogeneities which alter image intensities locally, an effect which can bias spatial preprocessing. To minimize these influences, image intensities were adjusted using the bias-correction procedure before spatial preprocessing.

### **2.3.6 Spatial preprocessing**

The assumption that a given voxel depicts the activity of the same part of the brain at every time point is incorrect if a subject moves, simply because voxel coordinates might refer to different locations in the brain before and after the movement. Therefore, head movement during scanning, even in the order of millimetres, can severely distort signal intensities. The realignment procedure corrects for such misattributions by aligning each EPI image with the first (or an alternative target) image in an experimental block through a rigid-body transformation (Andersson et al., 2001). Three translation and three rotation parameters are computed for each image such that the difference between this image and the first image in the sequence is minimized. However, not all artefacts can be removed by this procedure. Therefore, the estimated movement parameters are also later included in the design matrix as covariates.

The unwarp procedure corrects for scaling, shear as well as distortions and signal dropouts caused by susceptibility-induced field inhomogeneities. These geometric distortions are particularly prominent at air-tissue interface (e.g. orbitofrontal cortex and medial temporal lobes). As they are detrimental for functional activation maps they can interfere with accurate registration between an EPI image and a structural image. The geometric distortions were therefore directly measured using a fieldmap sequence recorded during the experiment and distortion correction was performed using the FieldMap toolbox implemented in SPM.

To align functional and anatomical images, the EPI images are coregistered with subject-specific anatomical scans so that task activations are overlaid on an individual's anatomical

scan. SPM performs this coregistration by computing a transformation matrix that maximizes the mutual information between the functional and structural images.

Functional and anatomical scans across subjects vary in size and shape, so consequently all images are transformed into a standard space using a normalization procedure to make inferences across the group. This procedure involves the warping of subject specific images with a 12-parameter affine transformation and a nonlinear transformation. The goal of this procedure is to minimise the differences between the Montreal Neurological Institute (MNI) reference brain and subject specific images. Normalized images can then be averaged across subjects, and results can be compared across studies and generalized to larger populations. The MNI space used here corresponds to an average of 152 MRI scans of right-handed healthy control subjects and is as such more representative of the general population than the often used Tairarach space which is based on a single subject.

A 3-dimensional Gaussian kernel with a full-width at half-maximum (FWHM) of 8 mm is then applied to the functional images. This spatial smoothing procedure reduces the noise in the data by averaging signal from neighbouring voxels. According to the matched filter theorem, signal-to-noise is optimal if the filter width matches the expected signal width. Furthermore, spatial smoothing increases the validity of the statistical analysis by making the error distribution more normal, and it reduces the effects of regional anatomical differences between subjects. Of course, spatial smoothing also has disadvantages such as a reduced spatial resolution.

### 2.3.7 Statistical analysis

fMRI studies typically aim at identifying brain regions or networks which respond to an experimental manipulation. The aim of the statistical analysis of fMRI data is therefore to find voxels in the brain whose neural signal significantly correlates with the measure of interest and cannot be explained by random signal changes alone.

The BOLD response is typically analysed using a linear regression model, or general linear model (Friston et al., 1994). Notably, in a standard univariate approach the regression is performed for each voxel, which are treated as independent units. The observed BOLD time series ( $Y$ ) is modelled as a linear combination of known explanatory variables or regressors specifying the onset of a condition, which are combined in a design matrix ( $X$ ), and a normally distributed error ( $\epsilon$ ). The amplitudes of the predictors  $\beta$  are then estimated such that the fit of the model's prediction to the data is maximized.

$$Y = \beta X + \varepsilon \quad (2.3)$$

The design matrix contains all experimental manipulations as well as effects of no-interest such as movement parameters, pulse and breathing, which are thought to contribute to the BOLD signal by causing noise. Furthermore, parametric modulators can be included if the amplitude of the neuronal response is thought to vary parametrically with trial-by-trial stimulus variations.

The signal at time  $t$  is then modelled as the convolution of the experimental stimulus function (i.e. the onset regressors, modelled as a boxcar for block designs and modelled as delta functions for event-related designs) and the hemodynamic response  $h(t)$  (Figure 2.5):

$$x(t) = (v * h)(t) \quad (2.4)$$

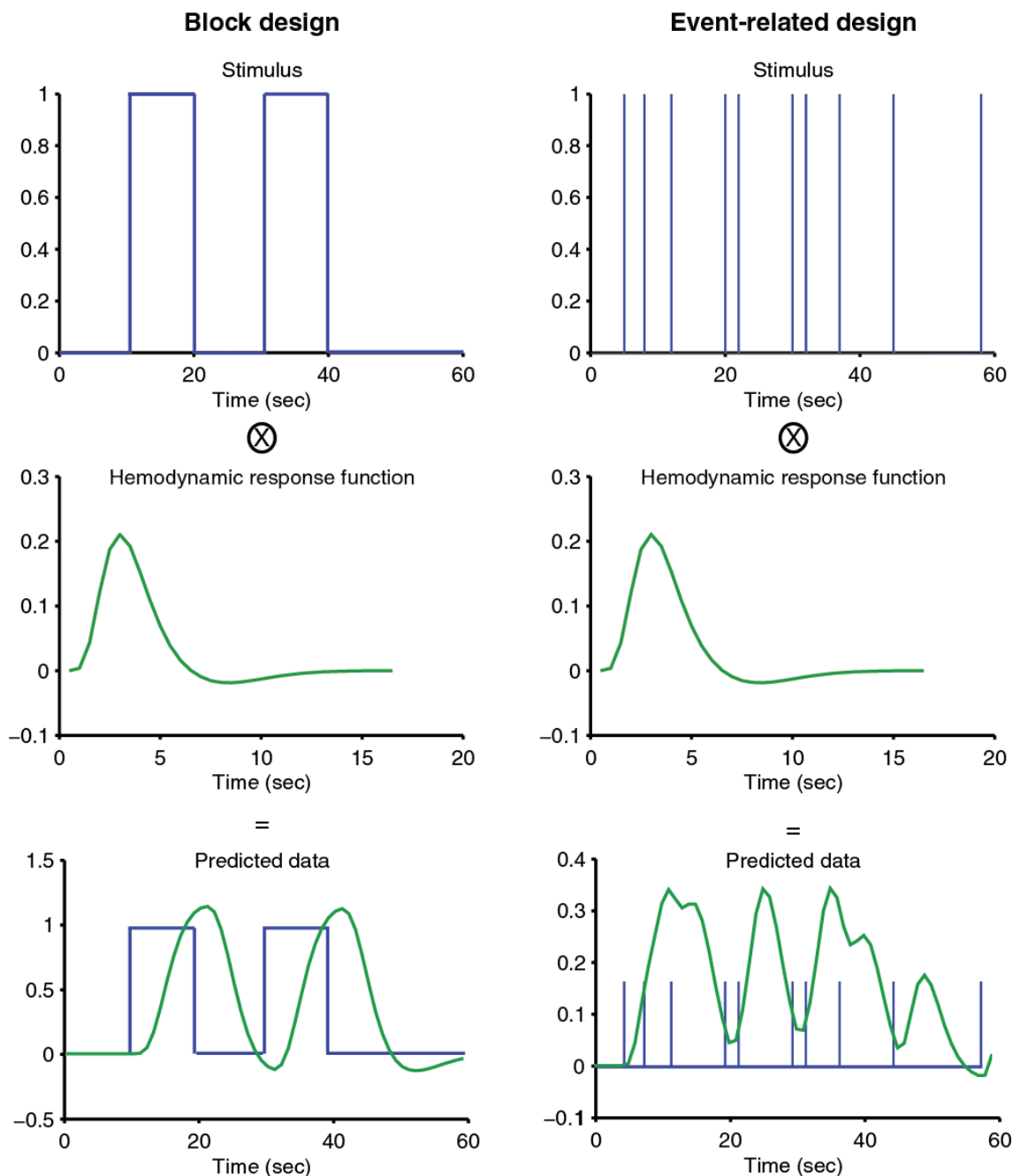
It is worth noting that the HRF might differ for different voxels as it depends on region-specific differences in vasculature and neural activity. To take such variability into account, temporal and dispersion derivatives can be added to the canonical basis function. In this thesis, convolution was performed with the canonical HRF alone.

### 2.3.8 Model estimation and statistical inference

The model estimation step aims at finding parameters  $\beta$  which provide the best fit to the data  $Y$  by minimizing the residual error. If  $\varepsilon$  is independent and identically distributed (i.i.d.), then  $\beta$  can be estimated using Ordinary Least Squares according to Eq. 2.5:

$$Y = X\beta + \varepsilon \leftrightarrow \beta = (X^T X)^{-1} X^T Y \quad (2.5)$$

Importantly, time-series data are in fact not independent. They are temporally correlated due to the shape of the HRF as well as due to the low-frequency noise contained within the fMRI signal (e.g. scanner drift). SPM adjusts for the resulting autocorrelations by adjusting the degrees of freedom.



**Figure 2.5 Convolution of condition regressors (top) with the HRF (middle) to predict neural activity (bottom).** Onsets are modelled as boxcars for block designs, and as delta functions for event-related designs.

The result of the model estimation is a number of beta images, which reflect the estimated parameters for each of the regressors in the design matrix. F and t test procedures can be used to test each voxel for significant activation in response to a stimulus. To test the null hypothesis that a linear combination of parameters is significant, a contrast vector is defined as follows:

$$cT\beta = c_1\beta_1 + c_2\beta_2 + \dots + c_n\beta_n \quad (2.6)$$

with  $\beta_1, \beta_2, \dots, \beta_n$  corresponding to the parameter estimates of interest. This test is performed using a T statistic, which is calculated according to the following formula:

$$T = \frac{c^T\beta}{\sqrt{\text{Var}(c^T\beta)}} \quad (2.7)$$

To simultaneously test whether any of a number of regressors explains any variance in the data, an F test is performed. An example scenario would include an F test across the movement parameters to test for areas correlating with subject movement (in any direction).

Both procedures create a statistical map indicating the T and F statistics for each voxel, respectively. To determine statistical significance, a threshold that carefully balances sensitivity and specificity needs to be chosen carefully. Furthermore, any result needs to be corrected for multiple comparisons. This correction accounts for the many false positives that are to be expected due to the very large number of tests that are performed across all voxels in the brain.

### 2.3.9 Group inferences

A second level random-effects analysis at the population level then allows for making inferences at the population level. Here, a one-sided t-test is performed to test whether a contrast is significant across subjects.

## 2.4 Tools for indexing neural computations at the meso-scale

Compared to other non-invasive recording techniques used to measure neural activity in the human brain, such as electroencephalography (EEG) or magnetoencephalography (MEG), fMRI allows for human brain activity to be measured at a high spatial resolution. This feature allows activity within the brain to be localized and enables the identification of brain regions with a specialized psychological function, such as areas specialized for face-, body- and place-related processing (Downing et al., 2001; Epstein and Kanwisher, 1998; Kanwisher et al., 1997), or emotion processing (Morris et al., 1996). More recently, model-based fMRI studies have identified brain regions performing particular computations, such

as reward prediction error (O’Doherty et al., 2003b) and value computations (Boorman et al., 2009; FitzGerald et al., 2009; Hayden et al., 2011). While these studies have been seminal for understanding functional specialization within the brain, neural processing in general cannot be fully characterized by a region’s average activation profile. A voxel contains more than  $10^5$  neurons, making it difficult to infer the computations performed by neuronal subpopulations from the average activation profile of a voxel. This is further complicated by a fundamental principle of neural information processing, namely the fact that information is typically encoded across a population of neurons in a distributed fashion (Averbeck et al., 2006; Haxby et al., 2001; Pouget et al., 2000). If we really want to understand how the brain encodes information relevant for behaviour, we need to understand how representations at the level of neural populations are transformed by relevant computations.

The spatio-temporal characteristics of a neural representation can be well characterized using large-scale electrophysiological recordings, which simultaneously provide high temporal and spatial resolution. Such recordings have yielded important insight into cortical information processing, and have contributed to our understanding of the population dynamics underlying motor responses (Churchland et al., 2012), choice (Mante et al., 2013) or memory consolidation (Hoffman and McNaughton, 2002). However, except under unusual circumstances such as pre-operative recordings in epilepsy patients (Ekstrom et al., 2003; Fell et al., 2001), direct recordings of neural activity are not feasible in the human brain due to their invasive nature. Yet, it is particularly important to get access to neural representations in the human brains. It is unclear whether the neural mechanisms underlying complex behaviours are preserved across species, and how the extensive training that is necessary for animals to perform complex tasks influences these mechanisms. Some higher cognitive functions, such as social behaviour, are likely not to be present to the same degree in species where single-cell activity is readily available. Furthermore, the contribution of brain areas that are unique to humans cannot be studied in primates and other mammals. In recent years, various attempts have therefore been made to refine the coarse signal available in fMRI and investigate neural responses at the meso-scale, i.e. at the level of neural populations. The most widely used techniques include fMRI adaptation or repetition suppression paradigms (Grill-Spector et al., 1999) and multi-variate pattern analysis (Haynes and Rees, 2006). fMRI adaptation relies on the suppression of neuronal responses observed upon repeated activation, a phenomenon which can be reliably observed across species, brain regions and conditions (Auzsztulewicz and Friston, 2016; Grill-Spector et al., 2006; Kregelberg et al., 2006; Larsson et al., 2015; Malach, 2012). fMRI multivariate pattern analysis (MVPA) takes

advantage of small biases in the distribution of neurons across neighbouring voxels, thereby providing a multivariate estimate of distributed activity (Haxby, 2012; Kriegeskorte et al., 2008; Norman et al., 2006; Quian Quiroga and Panzeri, 2009). In this thesis, fMRI adaptation paradigms have been used to investigate the neural plasticity associated with learning and consolidating new information.

#### **2.4.1 Repetition suppression as a tool for measuring the similarity of neural representations**

If the same stimulus is repeatedly presented in an experimental setting, then activity in neurons encoding any feature of the stimulus is suppressed, a phenomenon termed ‘repetition suppression’. Repetition suppression was first reported using electrophysiological recordings in primate inferotemporal (IT) cortex, where a suppression effect was observed in response to the repeated presentation of a light stimulus (Gross et al., 1967, 1969). Since its discovery, such a reduction in response with repetition has been observed across a wide range of brain areas and time scales (Albrecht et al., 1984; Baylis and Rolls, 1987; Blakemore and Campbell, 1969; Cohen-Kashi Malina et al., 2013; Miller et al., 1991, 1996; Riches et al., 1991) and even if other stimuli are interleaved (Li et al., 1993a). Repetition suppression is also present if the stimulus is task-irrelevant (Miller and Desimone, 1994), suggesting a functional role as an automatic short-term memory mechanism.

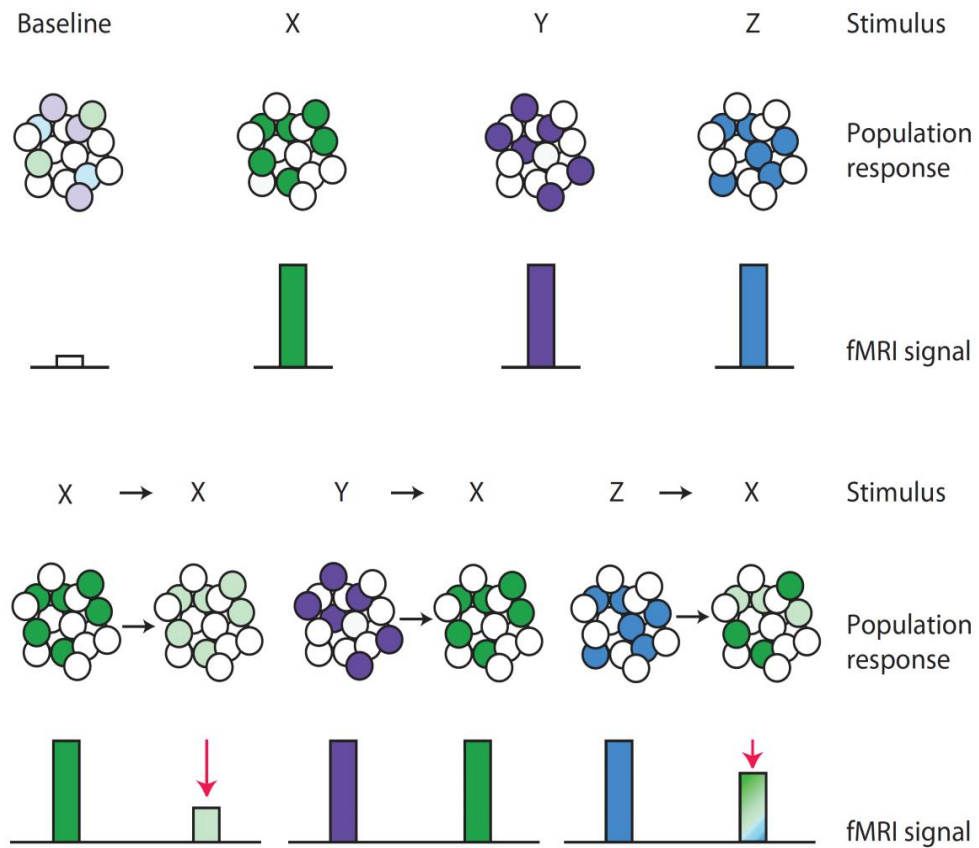
More recently, repetition suppression has been used to infer representational overlap between separate stimulus representations. In single-unit recordings in macaque area IT, a neuron that responds to both stimuli A and B shows a suppressed response not only if A is preceded by A, but also if A is preceded by B (Sawamura et al., 2006). This ‘cross-stimulus suppression’ effect is best explained by the similarity between the adaptor and test stimuli and by the strength of response to the adaptor stimulus (Baene and Vogels, 2010; Liu et al., 2009) rather than the neuron’s response to each stimulus per se (Piazza et al., 2007). This suggests that cross-stimulus adaptation scales with the amount of shared input between the two stimuli (Sawamura et al., 2006). Cross-stimulus suppression can therefore be utilized for probing a neuron’s selectivity and the representational overlap for different stimuli even, thereby providing access to both the nature and the relationship between different neural representations.

A similar reduction of stimulus-specific activity with repetition is widely observed in the human brain. Again, the phenomenon has been particularly well documented in visual areas,

where a repetition of the same visual object (Buckner et al., 1998) or face (Henson et al., 2000) leads to reductions in BOLD signal in lateral occipital and inferior temporal visual regions (Grill-Spector et al., 1999, 2006; Vuilleumier et al., 2002). Cross-stimulus suppression can also be observed in fMRI, even in a situation where the average response to different stimuli does not differ (Figure 2.6A). Neurons that contribute to the representations of both stimuli X and Z should show suppression to presentation of stimulus X preceded by Z (Figure 2.7B), resulting in a reduced BOLD signal relative to a situation where X is preceded by a stimulus Y which activates a non-overlapping neural representation (Figure 2.8B). This technique has proved useful in identifying different stages of object representation (Vuilleumier et al., 2002) or differentiating reward identity encoding from stimulus-reward association in medial OFC (Klein-Flügge et al., 2013b). Furthermore, cross-stimulus suppression can be used for indexing associative memories (Barron et al., 2016a), because they are formed by an increase in synaptic strength between the relevant cell assemblies, resulting in an increased overlap of the respective neural representations (Nabavi et al., 2014). Recently, it has even been used to track ongoing plasticity by measuring changes in association between a stimulus and reward representation over time (Boorman et al., 2016).

Furthermore, repetition suppression has been used to measure computations previously observed in animal models. One particularly striking example is the investigation of grid cells in the human brain. Grid cells are characterized by their hexagonally arranged firing fields which allow spatial knowledge to be organised into a map (Hafting et al., 2005). Remarkably, the phases of grid cells are aligned, which can be exploited using fMRI repetition suppression in the human brain. When subjects navigate through a virtual environment, the entorhinal cortex, mPFC, parietal cortex and temporal cortex show suppression as a function of running direction, modulated by running speed (Doeller et al., 2010). Crucially, this adaptation effect is selective to a running direction of  $60^\circ$ , consistent with six-fold rotational symmetry of the raw BOLD signal in the same brain regions and the predicted population response of grid cells. Together these studies illustrate how fMRI repetition suppression can be used to investigate the complex computations that underlie cognitive processes in the human brain.





**Figure 2.6 Illustration of the principle underlying fMRI adaptation.** The raw BOLD signal measured in conventional fMRI paradigms is invariant to the relationship between neural representations and instead provides only a measure of the mean activity of a population of neurons within a given voxel. In this example, the raw BOLD signal in response to stimuli X, Y and Z appears to be the same, because all representations recruit the same number of neurons. From the BOLD signal alone it is therefore not clear how the representations of stimuli X, Y and Z relate to one another. B In fMRI adaptation paradigms, the relationship between different stimulus representations X, Y and Z can be indirectly measured. If stimulus X is preceded by stimulus X (X-X), then the fMRI signal in areas encoding features particular to stimulus X are suppressed. If stimulus X is preceded by stimulus Y (Y-X), the response to X should not show any suppression, as the representations for X and Y are not overlapping. If X is preceded by Z (Z-X), the response in areas encoding the features that are shared between X and Z should show some suppression due to the overlapping representations of X and Z. Reproduced with permission from Barron et al. (2016b).

### 2.4.2 Biophysical mechanism underlying repetition suppression

Various attempts have been made at explaining the mechanisms underlying repetition suppression effects (Grill-Spector et al., 2006; Kohn, 2007). The ‘fatigue model’ assumes that repetition suppression is a consequence of low-level adaptation mechanisms resulting in a reduction of a neuron’s response. Here, the assumption is that the response amplitude, but not the response pattern, changes with repetition. Fatigue could in theory be due to reduced action potential firing caused by tonic hyperpolarization (Carandini and Ferster, 1997;

Sanchez-Vives and McCormick, 2000), or a synaptic depression mechanism which results in a reduced efficacy of synaptic inputs (Abbott et al., 1997; Carandini and Ferster, 1997). Recordings in macaque IT cortex suggest the latter is a more likely mechanism, because neurons responding equally to two stimuli A and B showed more suppression if A was preceded by A than if A was preceded by B (Sawamura et al., 2006). This demonstrates that repetition suppression, like LFP recordings and the BOLD signal, are likely to reflect shared local computations and integrated synaptic computations rather than action potential firing.

An alternative account of repetition suppression is the ‘sharpening model’, which assumes that information might be encoded more sparsely or efficiently with repetition, by recruiting fewer neurons and thereby sharpening a representation (Wiggs and Martin, 1998). Sharpening could be achieved by increased inhibitory input from lateral connections (Norman and O’Reilly, 2003), whereby neurons that are less selective cease to respond to a stimulus (Desimone, 1996). Under this model, repetition suppression not only alters the amplitude, but also the pattern of neuronal responses. While such effects are indeed observed in response to familiar stimuli and might therefore reflect learning, they fail to account for suppression in response to repeated novel stimuli. In contrast to the ‘fatigue model’, which predicts that neurons that respond most strongly to a stimulus show the largest amount of suppression, the ‘sharpening model’ predicts that the neurons most optimally tuned to a particular stimulus feature are the least affected by stimulus repetitions, which is inconsistent with the observations in single unit recordings (Li et al., 1993b; McMahan and Olson, 2007)..

The ‘facilitation model’ assumes that processing speed increases with repetition, due to shorter response latencies or firing durations (Grill-Spector et al., 2006). This model thus accounts for behavioural priming effects, whereby behavioural performance such as reaction time and accuracy improves in response to repeated stimulus exposure (Ferrand and Grainger, 1992; Schacter and Buckner, 1998; Tulving and Schacter, 1990). However, although repetition suppression and behavioural priming can be observed under similar conditions, the effects do not necessarily correlate (McMahan and Olson, 2007), suggesting that repetition suppression does not causally underlie behavioural priming.

Mechanistically, a ‘facilitation’ effect could occur due to synaptic potentiation or top-down modulations of activity. The ‘facilitation model’ is therefore also consistent with a predictive coding account of brain function. According to predictive coding, repetition suppression may be caused by top-down modulations from higher cognitive areas (Friston, 2005; Summerfield et al., 2008). The brain constantly tries to predict upcoming sensory

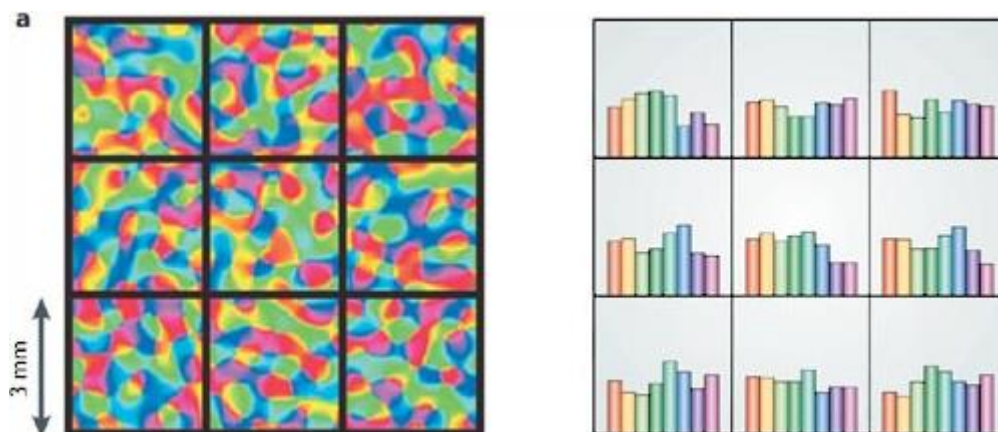
information. If a sensory input corresponds to the expected signal, top-down signals suppress a neuron's response. If, however, an expectation is violated or a novel stimulus is presented, the incoming sensory information is inconsistent with the predicted signal, and a prediction error signal is elicited. This account of repetition suppression makes the prediction that repetition suppression should be modulated by expectancy. This hypothesis could indeed be confirmed in an fMRI study where stimulus repetitions and expectation were modulated independently (Summerfield et al., 2008). Here, repetition suppression was stronger in response to stimuli that occurred with a higher frequency (75%) and were therefore less surprising than less frequent stimuli (25%). However, auditory evoked potentials measured using magnetoencephalography (MEG) in a situation where repetition and expectation were manipulated independently show that the two effects can be separated in time, with an early repetition suppression effect, and a later expectation suppression effect (Todorovic and Lange, 2012). This suggests that repetition suppression and expectation suppression need to be understood as two distinct processes, which may reflect prediction errors at different levels of the hierarchy (Todorovic and Lange, 2012).

### **2.4.3 Comparison of repetition suppression and multivariate pattern analysis methods**

An alternative approach for measuring representations are multi-variate pattern analysis (MVPA) methods. These methods again rely on the assumption that information is represented in the brain in a distributed fashion. However, while repetition suppression probes the representational similarity within voxels, in MVPA a representation is typically defined as the activity pattern across multiple voxels. Typically, a classifier such as a support vector machine is trained to discriminate between different multi-voxel patterns for different experimental conditions in a subset of the data, and the results are subsequently tested in an independent test data set. MVPA analyses can be performed within a particular region of interest (ROI), or across the whole brain in a searchlight analysis.

MVPA therefore relies on a fine-grained spatial structure across voxels within a brain region. Such a fine-grained spatial structure is particularly prominent in the visual cortex where neurons with a similar orientation selectivity are clustered in columns of several hundred micrometer width (Bartfeld and Grinvald, 1992; Obermayer and Blasdel, 1993). Even though the size of a voxel is much larger, visual features can still be successfully classified because orientation-selective columns are non-uniformly distributed (Figure 2.7). Similarly, neurons in IT cortex are organized in columns which share sensitivity to high level

features (Wang et al., 1996). As a consequence, MVPA has allowed for successfully decoding visual features from BOLD activity, such as the orientation (Haynes & Rees 2005a; Kamitani & Tong 2005), the perceived colour of a stimulus (Haynes & Rees 2005b) or object categories (Haxby et al., 2001).



**Figure 2.7 Orientation-selective cortical columns in the primary visual cortex.** Neurons sensitive to a particular edge orientation are clustered in cortical columns of approximately 500  $\mu\text{m}$  width (left). With a typical voxel size of 3mm, fMRI cannot resolve different cortical columns per se. However, computer simulations predict that the distribution of different object orientation sensitivities is not perfectly uniform across V1 (right), such that most voxels have a slightly higher number of columns sensitive to a particular orientation than expected by chance. Any presented bar should therefore cause slightly varying activity across voxels, which a sensitive machine learning algorithm can pick up. Reproduced with permission from Haynes and Rees (2006).

Other parts of the brain, however, do not display an architecture where neurons selective to a particular stimulus feature cluster within a particular cortical column. For example, response profiles of neurons in mPFC are very heterogeneous. Individual neurons display a nonlinear mixed selectivity to various features of a task (Rigotti et al., 2013), their tuning function can change throughout a task, and across the population the neural dynamics are very high-dimensional. Furthermore, mPFC receives inputs from a much wider array of cortical areas, as reflected cytoarchitecturally by long-ranging dendritic fields in mPFC (Jacobs, 2001). As a consequence, information encoded within mPFC is much more varied and less likely to display fine grained spatial structure than sensory areas. Nevertheless, MVPA has been successfully applied to classify information in mPFC and other prefrontal brain areas (Howard et al., 2015; Kahnt et al., 2010).

Representational similarity analysis (RSA) is a variant of MVPA. Like MVPA, RSA assumes that representations in the brain needs to be understood in terms of the multi-voxel activity patterns rather than bulk effects within a particular brain region. However, rather

than trying to decode information about stimuli from brain states, RSA focusses on the neural representation of stimuli and specifically characterizes the relationship between different stimulus representations, or the representational geometry within a brain region for each individual (Kriegeskorte and Kievit, 2013). Many computationally trivial transformations such as rotations could fundamentally alter the bulk activity within a voxel, without influencing the representational geometry across voxels. Furthermore, in individual subjects the relationship between stimulus representations can be highly invariant, even if the activity patterns for each stimulus varies greatly. Understanding a neural code within a brain region therefore requires an understanding of the relationship between stimulus representations in a high dimensional response pattern space (Kriegeskorte et al., 2008). RSA measures such a dissimilarity between stimulus representations and compares the resulting representational dissimilarity matrix to dissimilarity matrices derived from stimulus descriptions, behaviour or computational models. This approach has proven successful in characterizing low level visual processing (Hiramatsu et al., 2011), sensory and motor representations (Wiestler et al., 2011) and even memory (Xue et al., 2010).

In conclusion, repetition suppression, MVPA and RSA measure different aspects of the BOLD signal, and consequently differ in the sensitivity to particular features of the neural code (Drucker and Aguirre, 2009). Due to the architecture of sensory cortices, MVPA can be more sensitive than repetition suppression in sensory areas such as the visual cortex (Sapountzis et al., 2010). However, it is also important to note that repetition suppression and MVPA measures typically correlate (Sapountzis et al., 2010).

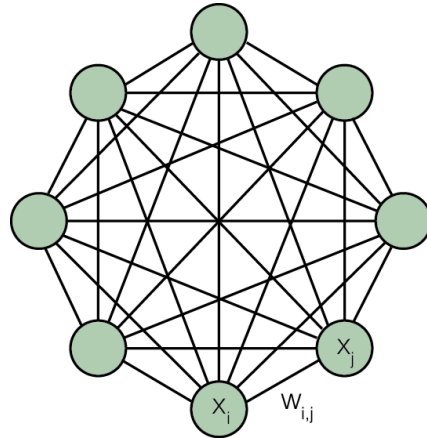
## 2.5 Hopfield networks as models of associative memory

In 1982, John Hopfield proposed a very simple neural network that is able to retrieve an entire stored pattern from partial information (Hopfield, 1982). This idea is reminiscent of Hebb's postulate that the entirety of a representation can be retrieved by activating a subpopulation of the cell assembly. The Hopfield network has therefore been hugely successful as a simple model of associative memory.

Hopfield networks can be described as complete, un-directed graphs with symmetric weights (Figure 2.8), where each node corresponds to a simple model of a neuron. The details of action potential timings of each model neuron are suppressed. Instead, activity  $x_i$  of each neuron in the network can be conceptualized as an instantaneous firing rate generating action

potentials stochastically.  $x_i$  is iteratively updated by computing the weighted sum of the inputs it receives from all the other neurons in the network, bounded by  $+1/-1$ :

$$x_i = \tanh \left( \sum_j w_{i,j} x_j \right) \quad (2.8)$$



**Figure 2.8 Schematic of a complete, undirected Hopfield network.**  $w_{i,j}$  denotes the connection strength between neurons  $x_i$  and  $x_j$

Memories are stored as  $n$ -dimensional patterns of activity  $\mu$  across all of the  $n$  neurons  $x_1, x_2, \dots, x_n$  and can be described as an  $n$ -dimensional location in state space. To store these memories, the network's connection weights need to be set according to a learning algorithm. The most common learning rule is a biologically plausible, auto-associative Hebbian learning rule, where the weight between two neurons  $w_{i,j}$  is given by the product of the pre- and post-synaptic activity  $x_i$  and  $x_j$ :

$$w_{i,j} = \frac{1}{N} \sum_{\mu} (x_i^{\mu} x_j^{\mu}), \text{ with } w_{i,i} = 0 \quad (2.9)$$

Applying this algorithm results in the memory patterns being stored in the weight matrix. Crucially, the stored memory patterns can then be considered stable attractor points in a state space of this neural system. Initializing the system with a cue which resembles one of the stored memory patterns  $\mu$  then corresponds to a situation where the network starts relatively nearer to the corresponding pattern in state space. For example, all neurons in the network can be initialized with a noisy version of memory pattern  $\mu$ :

$$x_i = \tanh(x_i^{\mu} + N(0,1)). \quad (2.10)$$

Conceptually, this initialization can be thought of as an activation of a small subpopulation of neurons forming a representation, in line with the Hebbian idea of a cell assembly outlined above (Hebb, 1949). The activity of each neuron is then iteratively updated according to the weight matrix. Here, the input to each neuron  $x_i$  is computed as the sum of the activity of all other neurons  $x_j$  multiplied by the synaptic weight  $w_{i,j}$  between neurons  $x_i$  and  $x_j$ , and bound between -1 and 1. This term was weighed with a factor  $dt$  and added to the current state of the neuron  $x_i$ , weighed by  $1 - dt$ :

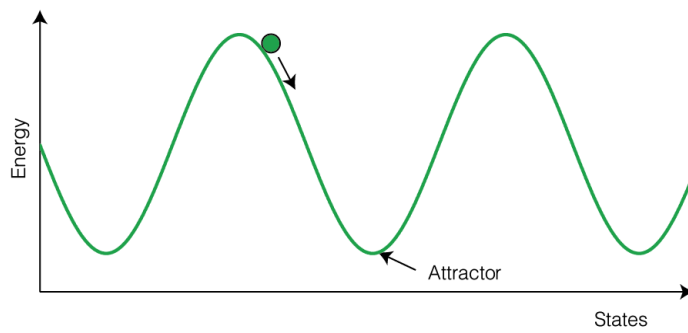
$$x_i = (1 - dt)x_i + dt * \tanh\left(\sum_j w_{i,j}x_j\right) \quad (2.11)$$

The activity levels of all neurons can be updated simultaneously (synchronous updating), or one unit is updated at a time (asynchronous updating). The network's state, which can be described as a state vector whose elements corresponds to the activity of each neuron in the network, then evolves in time in a way that reduces the value of a so called 'energy function'. The energy of a given neuron  $x_i$  in the Hopfield network is defined as:

$$E_i = -\frac{1}{2}\left(\sum_j w_{i,j}x_j\right)x_i \quad (2.12)$$

Importantly, any change in activity of neuron  $x_i$  will lead to a decrease in energy (Figure 2.9). If  $(\sum_j w_{i,j}x_j)$  is negative, then the change from  $x_i$  at time step  $t$  to  $x_i$  at time step  $t+1$  must be zero or negative (Eq. 2.13), such that  $\sum_j w_{i,j}x_j (x_i^{t+1} - x_i^t) \geq 0$ , and  $\Delta E \leq 0$ . If, however,  $(\sum_j w_{i,j}x_j)$  is positive, then  $x_i$  will become more positive, such that, again,  $\sum_j w_{i,j}x_j (x_i^{t+1} - x_i^t) \geq 0$  and  $\Delta E \leq 0$ . In summary,  $\Delta E_i$  is always zero or negative, such that the network must converge towards a local minimum. This local minimum corresponds to a stable attractor state, which is most similar to the cue.

$$\begin{aligned} \Delta E_i &= \left[-\frac{1}{2}\sum_j w_{i,j}x_j x_i^{t+1}\right] - \left[-\frac{1}{2}\sum_j w_{i,j}x_j x_i^t\right] \\ &= -\frac{1}{2}\sum_j w_{i,j}x_j (x_i^{t+1} - x_i^t) \end{aligned} \quad (2.14)$$



**Figure 2.9** Illustration of a one-dimensional energy surface.



### **3 LEARNING ABOUT THE PREFERENCES OF ANOTHER ALTERS SUBJECTIVE VALUATION IN DELAY DISCOUNTING**

\* This chapter is published in the following article:

Garvert MM, Moutoussis M, Kurth-Nelson Z, Behrens TEJ & Dolan RJ (2015).  
Learning-induced plasticity in medial prefrontal cortex predicts preference  
malleability. *Neuron* 85: 418-428

### 3.1 Abstract

Subjective preferences in decision-making situations vary widely across humans, and are often susceptible to social influence. Here, we used a Bayesian model of choice behaviour to investigate the mechanisms underlying social influence in intertemporal choice, where subjective preferences can be precisely quantified using a subject-specific discount rate parameter. We show that a subject's discount rate shifts towards the discount rate of a partner when the other's preferences are learned, irrespective of whether the other is human or a computer. Preference shifts do not arise if choices for the other can be based on a simple category-learning task, suggesting that simple stimulus- or action-reinforcement cannot account for a shift in subjective preference. Instead, preference shifts are a consequence of preference simulation, whereby the same discounting mechanism is employed when choosing for self and other. These findings provide evidence that intertemporal preferences, far from being a fixed trait, are modified in a manner that reveals a subtle but powerful influence of social learning mechanisms.

### 3.2 Introduction

Our perception, values and even memories are highly susceptible to the opinions, judgements and behaviour of others (Berns et al., 2010; Campbell-Meiklejohn et al., 2010; Edelson et al., 2011; Klucharev et al., 2009; Shestakova et al., 2013; Zaki et al., 2011). However, the mechanisms driving a change in our attitudes and behaviour as a consequence of social influence are not well understood (Cialdini and Goldstein, 2004). Psychological explanations provided for this phenomenon include a pursuit of acceptance of a social group (Deutsch and Gerard, 1955) or a pursuit of accuracy in situations where the other has additional information (Jetten et al., 2006). Here, we propose that in some instances social influence on subjective preference can be explained mechanistically as the consequence of a learning-induced plasticity in overlapping neural representations. In medial prefrontal cortex (mPFC), for example, the same neural circuitry performs value computations on behalf of oneself and on behalf of another person (Nicolle et al., 2012), and the same neural population performs self-referential as well as social value computations (Jenkins et al., 2008). In light of this, multiple value computations might be updated simultaneously if a learning-induced plasticity is introduced into this circuitry, resulting in a shift in preference. Here, we

investigate whether learning the preferences of another impacts on subjects' own preferences in a situation where subjective preferences can be precisely quantified, and where they vary substantially across individuals.

It has been well established for centuries that changes in the objectively measurable features of the environment do not translate linearly to changes in subjective perception. For example, a 'just noticeable difference' in weight between two objects is proportional to the absolute weight of the objects (Weber, 1834), and the relationship between subjective perception and objective features of the environment is logarithmic across many domains (Fechner, 1860). Similarly, in economic decision making the subjective value of a good does not necessarily relate linearly to objective features such as its amount (Park et al., 2011). A nonlinear mapping of subjective preferences onto objective values is also evident in situations where a reward can only be attained after a certain delay. In a standard intertemporal choice task, participants choose between a smaller, immediately available, reward and a larger, temporally delayed, reward. Empirically, participants value rewards less the longer they have to wait to obtain them. The steepness of this decrease in the subjective value of delayed options can be described by a hyperbolic function with a subject-specific discount rate parameter (Myerson and Green, 1995). This parameter varies widely across people, but is considered stable over time in an individual in the absence of an experimental manipulation (Kirby, 2009; Ohmura et al., 2006). Furthermore, discount rates are related to self-control abilities, and elevated in drug addiction (Kreek et al., 2005), problem gambling (Alessi and Petry, 2003), attention deficit/hyperactivity disorder (Winstanley et al., 2006) and other impulsivity disorders (Madden et al., 1997).

In this chapter, we investigate whether a Bayesian learning algorithm can be used to infer subjects' own discount rate preferences and their beliefs about the preferences of a (human or computer) partner from the choices subjects make in a delegated intertemporal choice paradigm. Critically, subjects own preferences are assessed before and after choosing on behalf of the partner to assess whether learning about another's preferences modulates subjects' own preferences. Performance in this 'mentalizing' condition is compared to a category-learning control experiment, where a decision for the other is based on a geometric depiction of the given options on the screen. This control experiment consisted of the same stimuli and actions, but the necessity to simulate another's preferences was removed.

### 3.3 Methods

#### 3.3.1 Participants

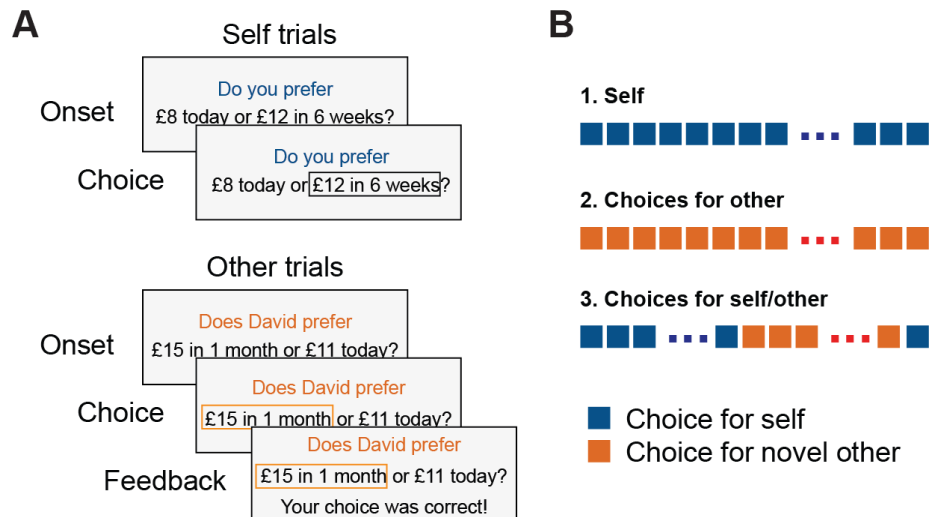
27 volunteers (mean age  $\pm$  std:  $23.6 \pm 3.7$ , 14 females) participated in the main behavioural experiment (human partner group) and 54 volunteers (mean age  $\pm$  std:  $23.0 \pm 7.9$ , 31 females) participated in the two control experiments. 27 were randomly assigned to a computer and a visual choice task condition, respectively. There were no significant differences in age ( $F_{2,78} = 0.71$ ,  $P = 0.49$ ) or gender ( $F_{2,78} = 1.04$ ,  $P = 0.36$ ) between the human partner, the computer partner, and the visual display group.

16 volunteers (mean age  $\pm$  std:  $22.4 \pm 2.9$ , 10 females) participated in the second 2-partner behavioural experiment. All subjects were neurologically and psychiatrically healthy. The study took place at the Wellcome Trust Centre for Neuroimaging in London, UK. The experimental procedure was approved by the University College London Hospitals Ethics Committee and written informed consent was obtained from all subjects.

#### 3.3.2 Human partner task

Pairs of gender-matched participants were introduced to each other as partners before the experiment and instructed simultaneously, but performed the task in separate rooms. Both subjects made a series of choices between a smaller amount paid on the same day and a larger amount paid later (Figure 3.1). The amounts varied between £1 and £20 and the delay was tomorrow, 1 week, 2 weeks, 4 weeks, 6 weeks, 2 months, or 3 months. The two options were presented simultaneously and the location of the immediate and delayed option on the screen was randomized. Subjects chose by pressing a button corresponding to the location of their preferred option on the screen without any time constraint.

In block 2, subjects were told that they are exposed to their partner's options from block 1 and were tasked to reproduce the partner's decisions. Choices were correct if they corresponded to the decision that would be preferred by a hyperbolic discounter with the discount rate used to generate the decisions (see below for details). Block 2 ended once subjects made 85% correct responses for their partner in a sliding window of 20 trials or after a maximum of 60 trials. In block 3, smaller blocks of ten trials of choosing for self, alternated with blocks of ten trials of choosing for the partner. Block 3 ended after a total of 200 trials.



**Figure 3.1 Experimental design.** **A** On each trial, subjects chose between an immediately available, smaller and a delayed, larger reward. On “self” trials, subjects considered the choice for themselves. On “other” trials, they made the choice on behalf of a partner, and feedback indicated whether their choice corresponded to the partner’s (simulated) choice. **B** Block 1 consisted of self choice trials alone, block 2 consisted of other choice trials alone and block 3 consisted of alternating short blocks of 10 choice trials per agent (self or other).

One of the outcomes chosen by the subject for themselves was randomly selected at the end of the experiment and transferred to their bank account after the respective delay. After finishing the experiment, subjects completed a debriefing questionnaire designed to assess the credibility of the experimental design. The questionnaire consisted of two questions:

- Did we communicate clearly that you would be confronted with the choices and evaluations your partners had been exposed to before, and the feedback you received was based on the decisions your partners had made?
- Was it clear to you that your payment depends only on the decisions you made for yourself, and that the same applied to your partners’ payment

All subjects believed in our experimental manipulation (Figure 3.2). Subjects were also instructed that the choices they made for the other were not communicated to the partner and did not have any consequences for either subject.

### 3.3.3 Estimation of discount rates using Bayesian modelling

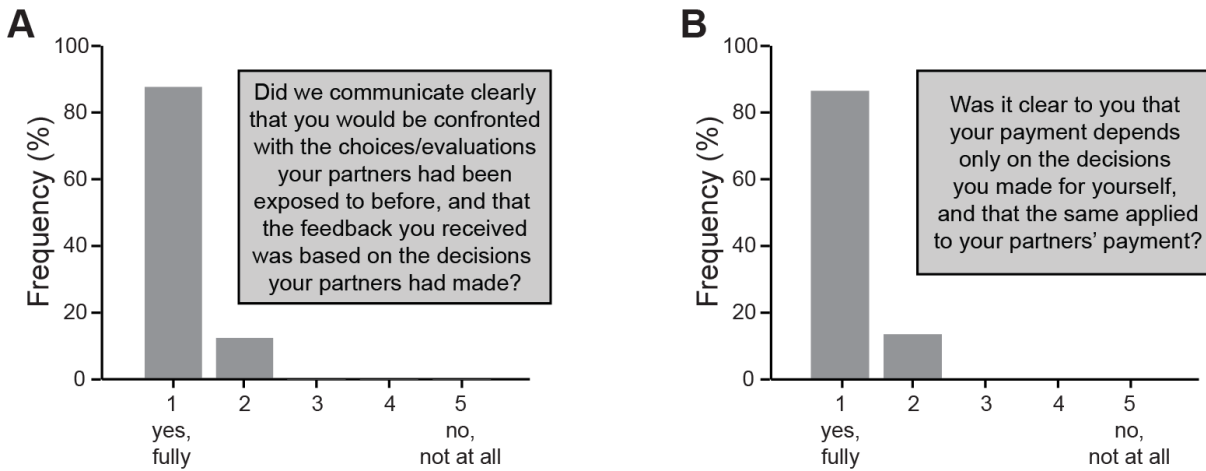
I characterized subjective preferences in the different blocks by estimating subject-specific discount rates  $k$ , which quantify the devaluation of future rewards according to a hyperbolic discounting model (Rachlin et al., 1991):

$$V = \frac{M}{1 + kD} \quad (3.1)$$

Here,  $V$  is the subjective value of an option,  $M$  is the magnitude, and  $D$  is the delay. When  $k = 0$ , subjects do not discount future rewards and base their valuation of an option purely on its magnitude, without regard for delay. As  $k$  grows, subjects discount future rewards more and more steeply. Since the delay of the smaller option was always 0 (today), the subjective value of the smaller option ( $V_{SS}$ ) always corresponded to its magnitude. Choice can then be modelled using a softmax function, whereby the difference in value between the smaller/sooner option and the larger/later option is translated into a choice probability according to the following equation:

$$P(y|k, \beta) = \frac{1}{1 + e^{-\beta(V_{LL} - V_{SS})}} \quad (3.2)$$

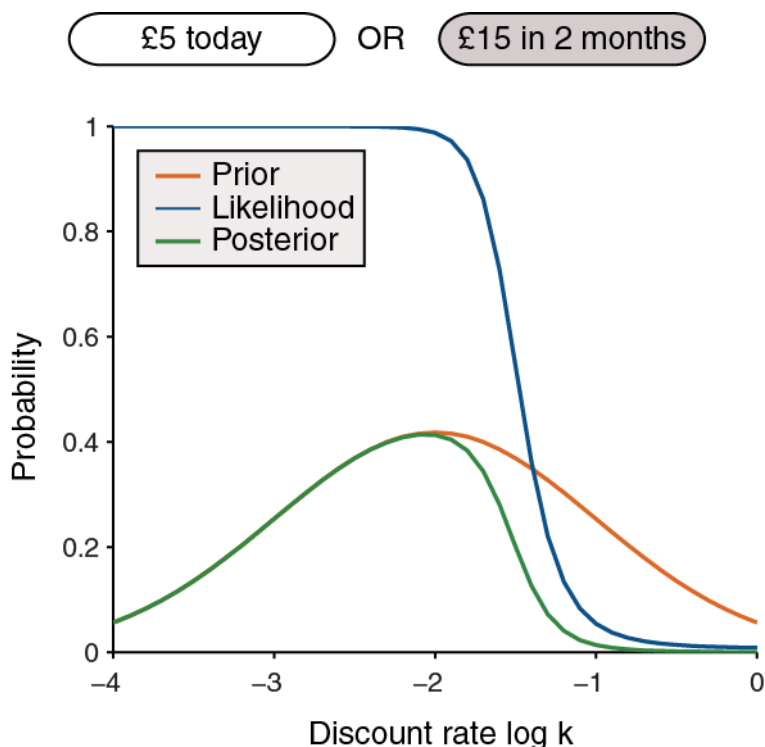
Here,  $\beta$  is a subject-specific inverse temperature parameter that characterizes non-systematic deviations around the indifference point.



**Figure 3.2 Results of debriefing questionnaire.** Illustrated is the percentage of subjects answering 1 (yes, fully) to 5 (no, not at all) in response to the following questions: **A** “Did we communicate clearly that you would be confronted with the choices and evaluations your partners had been exposed to before, and the feedback you received was based on the decisions your partners had made?” and **B** “Was it clear to you that your payment depends only on the decisions you made for yourself, and that the same applied to your partners’ payment?”

I estimated subject-specific discount rates  $k$  as well as individual temperature parameters  $\beta$  from subjects’ choices in the different blocks using Bayes rule. According to a Bayesian inference framework, a prior belief distribution about the state of the world  $P(k, \beta)$  (in this

case: a belief about a subject's discount rate  $k$  and inverse temperature parameter  $\beta$ ) is updated every time new data is encountered, i.e. when subjects make a new decision. The prior probability  $P(k, \beta)$  over discount rates  $k$  and temperature parameters  $\beta$  was initially uniform within a range of  $\log k = [-4; 0]$  and  $\log \beta = [-1; 1]$ .



**Figure 3.3 Schematic visualization of discount rate estimation.** Based on a subject's decision (here: £15 in 2 months) the likelihood for each possible discount rate  $k$  is computed. The likelihood is then multiplied with the prior to yield the posterior. The posterior is used as the prior on the next trial. Note the second parameter  $\beta$  is neglected in this schematic for simplicity.

The likelihood that a given choice (smaller/sooner or larger/later) would be made given all possible combinations of  $k$  and  $\beta$  was computed according to Eq. (3.2). A posterior distribution  $P(k, \beta|y)$  can then be inferred by updating the prior distribution  $P(k, \beta)$  with the observed data  $y$  according to Bayes' rule (Figure 3.3):

$$P(k, \beta|y) = \frac{P(y|k, \beta) * P(k, \beta)}{P(y)} \tag{3.3}$$

A trial-by-trial discount rate estimate could then be computed as the weighted sum of the posterior distribution over all discount rates  $k$  according to the following equation:

$$k_t = \frac{\sum(k * \text{posterior})}{\sum \text{posterior}} \quad (3.4)$$

The posterior distribution on trial  $t$  was then used as the prior distribution on trial  $t+1$ , providing us with trial-by-trial estimate of subjects' discount rates and temperature parameters.

### 3.3.4 Simulation of the other's choices

To generate feedback for the confederate's choices, we simulated a partner with a discount rate that differed from the subject's own baseline discount rate by 1, i.e.  $\log k_{\text{other}} = \log k_{\text{self,block1}} \pm 1$ . Choices were correct if they corresponded to the decision that would be preferred by a hyperbolic discounter with this discount rate. Importantly, the simulated partner's choices were probabilistic, as the other's subjective value Eq. (3.1) was translated to a choice probability with a softmax function (temperature parameter  $\beta = 1$ ) according to Eq. (3.2).

### 3.3.5 Behavioural modelling of a subject's belief about the other's preferences

We applied a similar Bayesian model to track a subject's beliefs about their partner's preferences over trials. We assumed that participants had a prior belief about the partners' discount rate, which can be characterized by a log-normal distribution  $\log N(k_0, \sigma_0^2)$  and will be updated whenever subjects' received feedback about the actual choice of the other. We aimed at identifying that combination of initial  $k_0$  and standard deviation  $\sigma_0^2$  that best fit the subject's choices behaviour on behalf of the other in the task.

To this end, we estimated the probability of observing the choice behaviour that participants displayed for all possible combinations of  $k_0$  and  $\sigma_0^2$ . For each combination, we calculated the initial expected discount rate EV for the partner as:

$$k_0 = \frac{\sum(k \times \text{prior})}{\sum \text{prior}}, \text{ with prior} \sim \log N(k_0, \sigma_0^2) \quad (3.5)$$

This initial discount rate could be used to compute the likelihood of observing the choice that participants executed ( $\text{likelihood}_{\text{ex}}, P_{\text{ex}}(y_{\text{ex}}|k, \sigma^2, \beta)$ ) as well as the choice that they observed when receiving feedback ( $\text{likelihood}_{\text{obs}}, P_{\text{obs}}(y_{\text{obs}}|k, \sigma^2, \beta)$ ), these are the same if



the participant chooses correctly for the partner) for all possible discount rates  $\log k_0$  ranging from -4 to 0 and  $\log \beta$  according to Eq. (3.2). Based on the thus determined likelihoods, we calculated the posterior probabilities according to Bayes' rule for the executed and the observed choice:

$$P_{\text{ex}}(k, \sigma^2, \beta | y_{\text{ex}}) = \frac{P_{\text{ex}}(y_{\text{ex}} | k, \sigma^2, \beta) * P(k, \sigma^2, \beta)}{P(y)} \quad (3.6)$$

$$P_{\text{obs}}(k, \sigma^2, \beta | y_{\text{obs}}) = \frac{P_{\text{obs}}(y_{\text{obs}} | k, \sigma^2, \beta) * P(k, \sigma^2, \beta)}{P(y)} \quad (3.7)$$

Based on the assumption that subjects updated their beliefs according to the decisions they observed during feedback, the posterior<sub>obs</sub> distribution on trial t was then used as the prior distribution on trial t+1. The log evidence for the model and the parameters was estimated on each trial according to:

$$E_t = \log\left(\sum \text{posterior}_{\text{ex}} * \text{likelihood}_{\text{ex}}\right) \quad (3.8)$$

The goodness of fit across trials for the given parameter combination can then be estimated as:

$$Q(\log k_0, \sigma_0^2, \beta) = \sum E \quad (3.9)$$

After estimating this for all parameter combinations, we summed the exponential of this distribution over the irrelevant dimensions to determine the parameter combination  $(k_0, \sigma_0^2, \beta)$  with the highest evidence. This particular parameter combination was used to define the participant's prior belief about the partner's discount rate. To estimate changes in belief over trials, the described procedure was reiterated starting from a prior defined according to the best fitting parameter combination  $(k_0, \sigma_0^2, \beta)$  to estimate the change in belief over trials.

The trial-by-trial estimate of subjects' own discount rate also allowed for quantifying the surprise that subjects experienced when observing their partner's choice on each trial. This surprise was simply computed by subtracting the subject's probability to make the same choice herself from 1.

$$\text{surprise} = 1 - \frac{1}{1 + e^{-\beta(V_{LL} - V_{SS})}} \quad (3.10)$$

Note,  $V_{SS}$  and  $V_{LL}$  are computed based on a trial-by-trial estimate of subjects' own discount rate based on Eq. (3.1).

Furthermore, on each trial a prediction error can be estimated by calculating the discrepancy between the estimate of the participants' belief  $EV_{ex}$  about the partner's discount rate based on his choice for the partner, and the updated belief  $EV_{obs}$  after receiving the feedback:

$$PE_t = \frac{\sum(k \times \text{posterior}_{ex})}{\sum \text{posterior}_{ex}} - \frac{\sum(k \times \text{posterior}_{obs})}{\sum \text{posterior}_{obs}} \quad (3.11)$$

On trials on which the participant made the correct choice, the two terms, and consequently the prediction error, are 0.

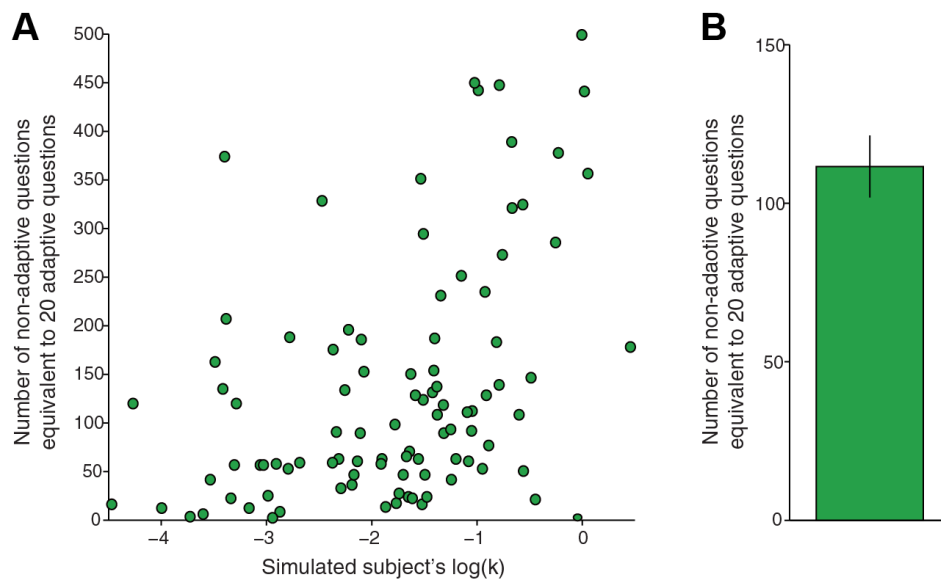
### 3.3.6 Optimization of choice pairs

In order to accurately estimate subjects' shift in discount rate, an efficient and precise estimate of subjects' discount rates was needed. To optimize choice pairs for this purpose, two option generation methods on choice trials for self were alternated. In method 1, first generated all possible pairs of amounts and delays and selected a subset of  $n/2$  trials ( $n$  = total number of trials in a block) that best matched the indifference points of  $n/2$  hypothetical subjects whose discount rates  $\log k$  were evenly distributed between  $[-4:0]$  in  $\log_{10}$  space (Nicolle et al., 2012). This procedure allowed for an efficient, but relatively imprecise, estimate of subjects' discount rates.

To increase the precision of our estimate of subjects' discount rates, I alternated the generated trials with choices generated according to a second method, which was adaptive in nature and based on the same Bayesian framework outlined above for estimating subjects' discount rates on a trial-by-trial basis. Here, the population distribution of  $\log k$  with a mean of -2 and a standard deviation of 1 was taken as a prior belief about an individual's  $\log k$ . Every time the subject made a decision, this belief distribution about their  $\log k$  was updated using the above described Bayes rule. Questions were then generated to specifically probe our current estimate of subjects' indifference point (where both options are equally preferred). The thus generated choices were more informative about the subjects' exact

discount rate than many of the other choice trials, such that this procedure gave us a more precise estimate of the subjects' discount rate. I validated the adaptive Bayesian method against the standard non-adaptive method and found that the adaptive method produced similar results as the non-adaptive method using fewer trials (Figure 3.4). The absolute difference between the actual discount rate and the estimated discount rate did not differ between the two methods ( $t_{1,94} = 0.26$ ,  $P = 0.79$ ). However, data from both methods were included in the final analysis to maximize power.

Choice pairs on “other” trials, were selected as a set of trials that best matched the indifference points of 60 (block 2) vs. 100 (block 3) hypothetical subjects whose discount rates were evenly distributed across the range centred on the other's discount rate [ $\log k_{\text{other}} - 1$ ;  $\log k_{\text{other}} + 1$ ]. This ensured that the number of immediate and delayed choices the subject made for the partner was approximately equal.



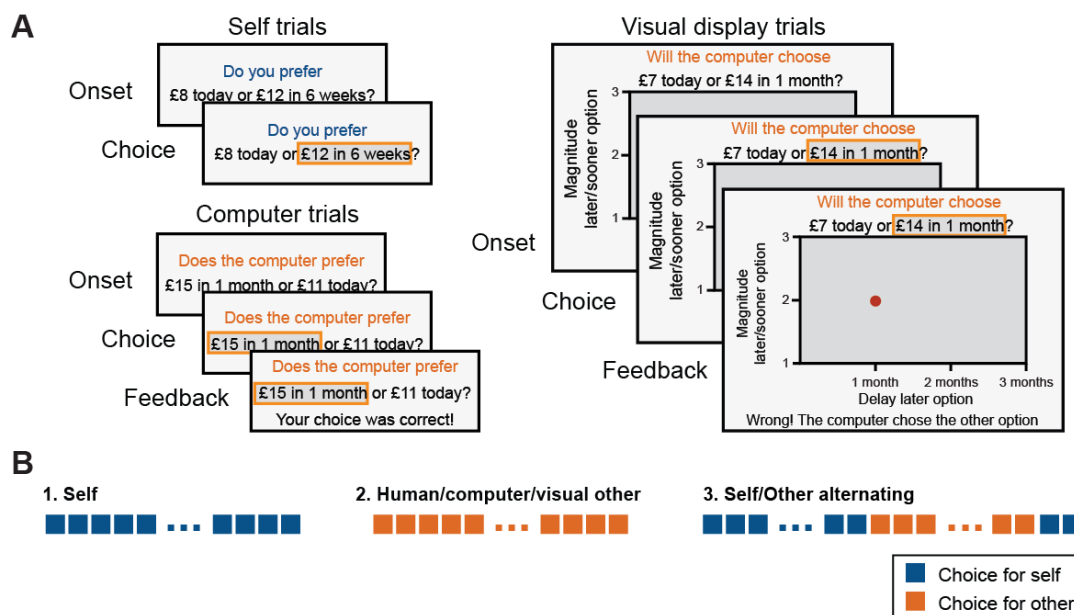
**Figure 3.4 Validation of adaptive method.** To compare the estimate obtained from the adaptive method versus the non-adaptive method, I first computed the variance of our discount rate estimate after 20 trials that were generated using the adaptive method. In a second step, I computed the variance of the discount rate estimated for choices that were generated using the non-adaptive method on a trial-by-trial basis. This iteration was terminated when the variance of the estimate obtained from the non-adaptive method was equal or smaller than the variance of the estimate computed based on the adaptive method. **A** Across all simulated discount rates more non-adaptive trials were required to obtain an estimate that was equally precise as the estimate obtained from the adaptive method. **B** The difference in number of trials needed was significant across simulated subjects ( $t = 9.26$ ,  $P < 0.001$ ).

### 3.3.7 Computer partner and visual display conditions

Subjects in the computer partner condition were told that a computer programme was trained to make decisions according to a specific strategy, while all other experimental settings were the same as in the human partner group. In actual fact, choices were generated in the same way as in the human partner condition.

Subjects assigned to the visual display condition learned a discount rate without engaging in any form of mentalizing. Instead, subjects were presented with a geometric depiction of a given choice on the screen (Figure 3.5, right) where the x-axis of the rectangle represented the delay of the delayed option and the y-axis represented the ratio of magnitudes for the delayed and the immediate options ( $M_{LL}/M_{SS}$ ). Subjects were told that the computer was programmed to choose one of the two options according to the location of the dot relative to an iso-probability line, which they had to learn based on the feedback they received after each choice.

In fact, choice in all three versions of the experiment were generated according to the above-described method.



**Figure 3.5 Design of computer partner and visual display control experiments.** **A** On self trials, subjects chose for themselves between an amount of money available on the same day and a larger amount of money available after a delay. On computer partner trials, subjects made these kinds of choices on behalf of a computer (left). On visual display trials, the choice pair was presented on a 2D grid. Subjects were instructed to choose according to their belief about the orientation of an imaginary isoprobability line. After each computer and visual display trial, feedback indicated whether a choice was correct. **B** Experimental structure: Block 1 consisted of self choice trials alone, block 2 consisted of other choice trials alone and block 3 consisted of alternating short blocks of 10 choice trials per agent (self or other). Block 2 terminated after 17 correct choices for the confederate within a sliding window of 20 consecutive trials or a maximum of 60 trials.

The orientation of the iso-probability line that participants had to learn was determined by the discount rate  $k$ , with larger discount rates corresponding to a steeper line. This follows from comparing choices for which the value of the immediate and the delayed option are equal:

$$V_{SS} = V_{LL} \Leftrightarrow M_{SS} = \frac{M_{LL}}{1+kD_{LL}} \Leftrightarrow \frac{M_{LL}}{M_{SS}} = kD_{LL} + 1 \quad (3.12)$$

Participants were instructed that dots above the line correspond to choosing the delayed, and dots below the line correspond to choosing the immediate option. Again, choice was translated into probabilities with a softmax function according to Eq. (3.2) such that choices were noisy. This was communicated to the participants.

### 3.3.8 Discount rate shift analyses

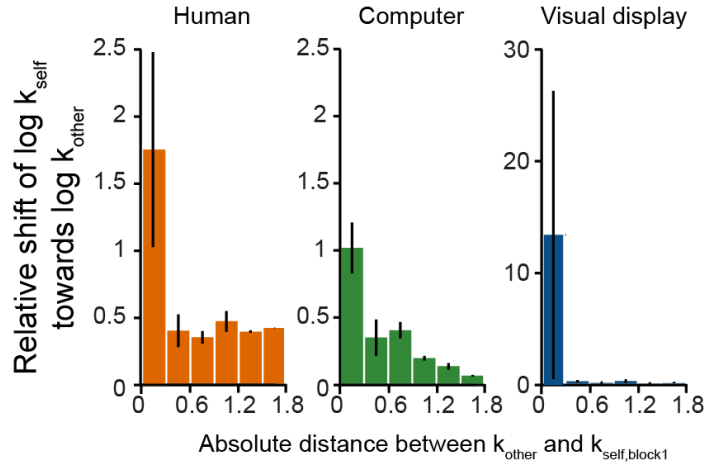
In all three versions of the experiment, computed subjects' shift in discount rate towards the novel other as follows:

$$\text{shift}_{\text{self} \rightarrow \text{novel}} = \frac{\log k_{\text{self}, \text{block 3}} - \log k_{\text{self}, \text{block 1}}}{\log k_{\text{novel}, \text{block 2}} - \log k_{\text{self}, \text{block 1}}} \quad (3.13)$$

Some participants estimated their confederate's discount rate to be very similar to their own. As a result even small shifts of subjects' own discount rate were substantially inflated when comparing it to the distance  $\log k_{\text{novel}, \text{block 2}} - \log k_{\text{self}, \text{block 1}}$ . Therefore, participants with an absolute distance  $|\log k_{\text{novel}, \text{block 2}} - \log k_{\text{self}, \text{block 1}}| < 0.3$  were excluded from all shift analyses. This value seemed to constitute a tipping point in all three experimental groups with strongly overrated shifts in discount rate for subjects with  $|\log k_{\text{novel}, \text{block 2}} - \log k_{\text{self}, \text{block 1}}| \leq 0.3$  (Figure 3.6). According to this criterion, I excluded 5 subjects in the human partner condition, 3 subjects in the computer partner condition and 2 participants in the visual display condition.

To test the robustness of our procedure to the particular threshold I chose, I also examined the results for a threshold of 0.5. This excluded 12 subjects in the human partner condition, 4 subjects in the computer partner condition and 2 subjects in the visual display condition. Importantly, I found that this threshold did not impact on the discount rate shift results, subjects in the human and computer, but not the visual display group to still shifted

towards the preferences of their partners: ( $t_{14} = 2.76$ ,  $P = 0.02$ ,  $t_{22} = 3.89$ ,  $P = 0.001$  and  $t_{24} = 0.61$ ,  $P = 0.5$ , respectively) with a significant difference between groups ( $F_{2,60} = 3.7$ ,  $P = 0.03$ ). Note that all discount rate analyses in this thesis were performed in  $\log_{10}$  space, transforming typical discount rates of  $[0.0001 - 0]$  to the range  $[-4 - 0]$ .



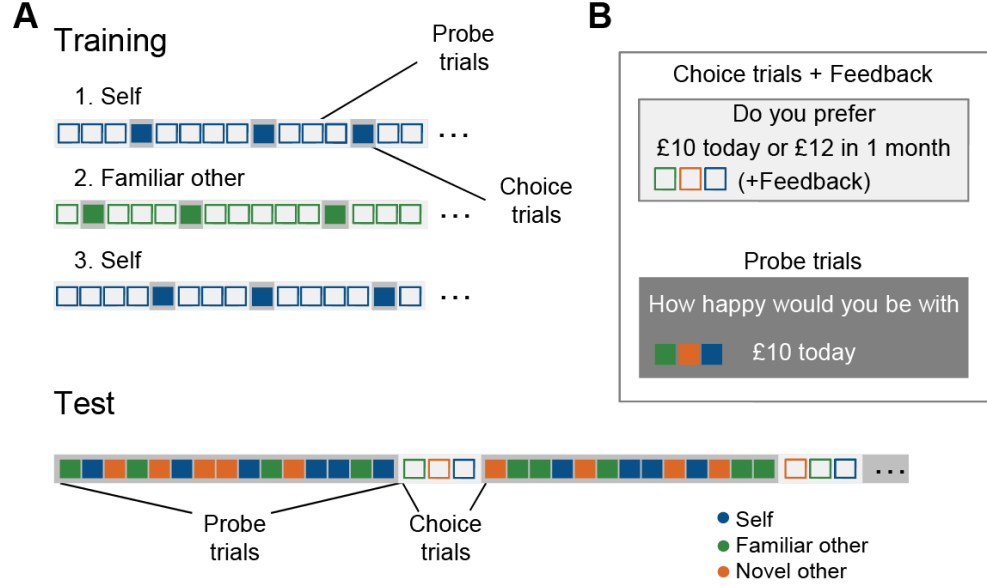
**Figure 3.6 Discount rate shift binned according to the distance between  $k_{\text{self}}$  and  $k_{\text{other}}$ .** Discount rate shift ( $\text{shift} = \frac{\log k_{\text{self,block3}} - \log k_{\text{self,block1}}}{\log k_{\text{other,block2}} - \log k_{\text{self,block1}}}$ ) binned according to the distance between  $|\log k_{\text{other,block2}} - \log k_{\text{self,block1}}|$ . As shift estimates were inflated for  $|\log k_{\text{other,block2}} - \log k_{\text{self,block1}}| \leq 0.3$ , subjects who estimated the other’s discount rate to be within that range were excluded from all discount rate shift analyses.

### 3.3.9 Two-partner behavioural experiment

In the 2-partner version of the experiment, a set of probe trials was added to the experiment (Figure 3.7B). The combinations of amount and delay on probe trials were drawn randomly from the same set as the options presented in choice trials. Participants were instructed to choose according to their own preferences (“self” trials) or according to the choice their partners had made when they had participated in the same experiment previously (“other” trials).

Subjects learnt the preferences of a second partner (“familiar other”) during training (Figure 3.7A). In this session, participants performed one block of choices and evaluations for self before and after a block of choices for the partner. Each pre-training block contained 48 choice trials as well as 16 randomly interleaved probe trials to familiarize subjects with this trial type. Each of the three experimental test blocks consisted of 197 probe trials for self, novel other and familiar other and 16 short interleaved blocks of one choice trial per agent (Figure 3.7A, bottom).

Again, participants were informed that one of the choices for self in the experiment would be randomly selected and the amount would be transferred to the participant's bank account after the appropriate delay. Decisions in probe trials did not influence the pay-out.



**Figure 3.7 Experimental design of the two-partner version of the experiment.** **A** During training, blocks 1 and 3 consisted of self choice and probe trials, and block 2 consisted of ‘familiar other’ choice and probe trials. During test, subjects chose and evaluated for themselves, for the familiar other and for a novel other. The experiment was divided into three experimental blocks with probe trials the predominant type in all blocks. **B** Trial types. On choice trials, subjects chose between an immediately available, smaller and a delayed, larger reward. On “self” trials, subjects considered the choice for themselves. On “other” trials, they made the choice on behalf of a partner, and feedback indicated whether their choice corresponded to the partner’s (simulated) choice. On probe trials subjects indicated on a four-item scale how happy they themselves or one of their partners would be with the presented option. These trials are not relevant here, but they are analysed in the fMRI version of the experiment (Chapter 4).

In contrast to the behavioural experiment and the training, subjects learned about the novel other’s discount rate while we assessed their own discount rate. To make sure that we captured a potential shift in discount rate in this scenario, we excluded the first third of all choice trials subjects performed in the test phase of the experiment. The relative shift effects reported in Figure 3.13 were then calculated as:

$$\text{shift}_{\text{self} \rightarrow \text{fam}, \text{train}} = \frac{\log k_{\text{self}, \text{training\_block 3}} - \log k_{\text{self}, \text{training\_block 1}}}{\log k_{\text{familiar}, \text{training\_block 2}} - \log k_{\text{self}, \text{training\_block 1}}} \quad (3.14)$$

$$\text{shift}_{\text{self} \rightarrow \text{fam}, \text{test}} = \frac{\log k_{\text{self}, \text{test}} - \log k_{\text{self}, \text{training\_block 3}}}{\log k_{\text{familiar}, \text{test}} - \log k_{\text{self}, \text{training\_block 3}}} \quad (3.15)$$

$$\text{shift}_{\text{self} \rightarrow \text{novel}, \text{test}} = \frac{\log k_{\text{self}, \text{test}} - \log k_{\text{self}, \text{training\_block3}}}{\log k_{\text{novel}, \text{test}} - \log k_{\text{self}, \text{training\_block3}}} \quad (3.16)$$

$$\text{shift}_{\text{fam} \rightarrow \text{novel}, \text{test}} = \frac{\log k_{\text{familiar}, \text{test}} - \log k_{\text{familiar}, \text{training\_block2}}}{\log k_{\text{novel}, \text{test}} - \log k_{\text{familiar}, \text{training\_block2}}} \quad (3.17)$$

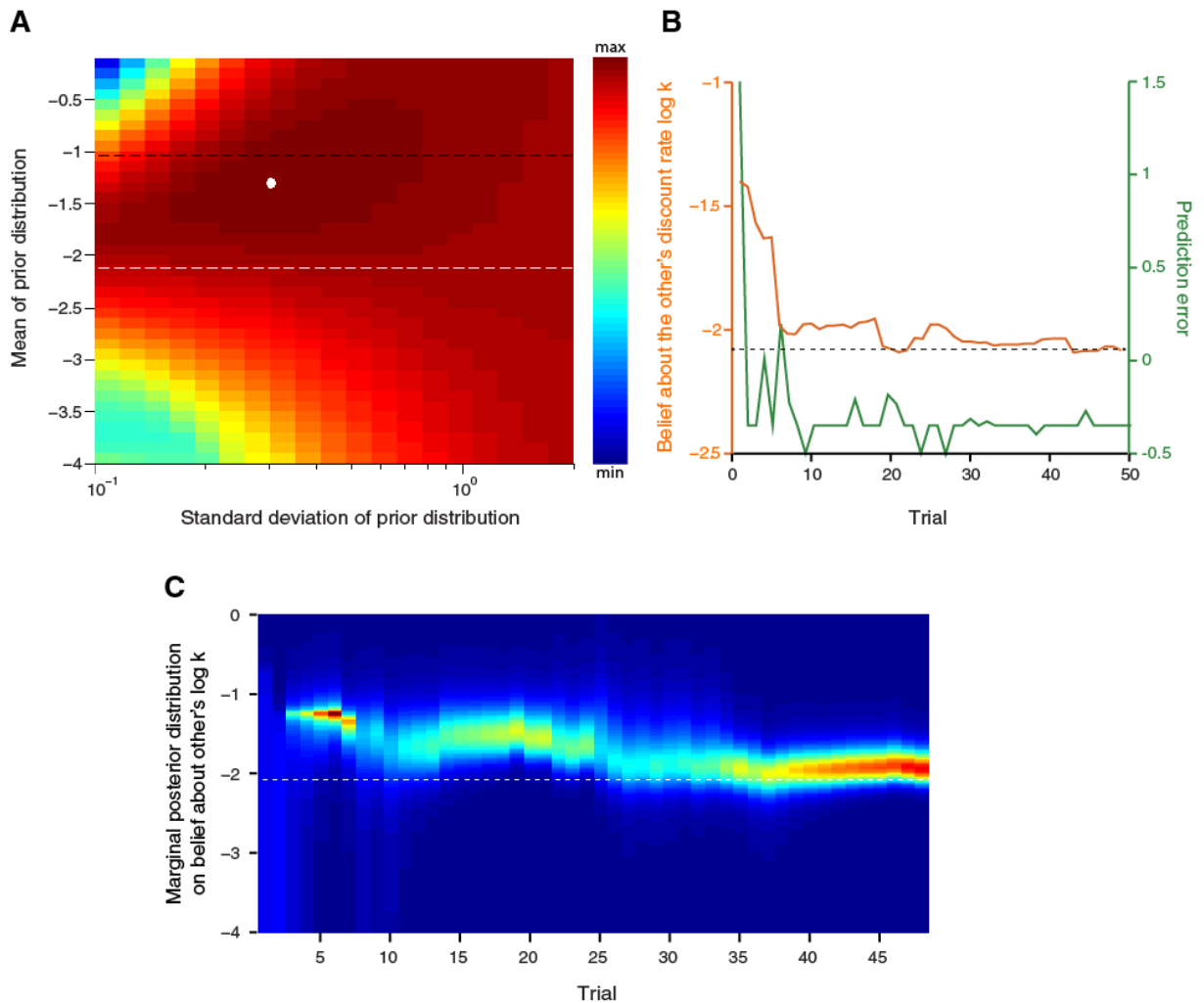
In line with the procedure outlined in 3.3.8, subjects for whom the denominator was smaller than 0.3 (1 subject for  $\text{shift}_{\text{self} \rightarrow \text{fam}, \text{training}}$ , 3 subjects for  $\text{shift}_{\text{self} \rightarrow \text{novel}, \text{test}}$ , 4 subjects for  $\text{shift}_{\text{self} \rightarrow \text{familiar}, \text{test}}$  and 2 subjects for  $\text{shift}_{\text{fam} \rightarrow \text{novel}, \text{test}}$ ) were excluded from the analyses.

## 3.4 Results

### 3.4.1 Subjects' behaviour can be modelled using a Bayesian learner

To examine whether learning about the preferences of another agent impacts on subjective inter-temporal preferences we tested 27 subjects on a standard inter-temporal choice task both before, and after, performing the identical task on behalf of a partner (Figure 3.1). As in the standard format, subjects deciding for themselves chose between an immediately available smaller reward and a delayed larger reward. The degree to which delay diminishes the value of a reward was then quantified by a discount rate, computed from each subject's actual choices both before and after the experimental manipulation. The latter involved a context whereby subjects performed the very same task but now chose the option they inferred a confederate would prefer. After each trial they were given feedback about the choice the confederate had actually made, such that they could learn to simulate these choices in future trials. Subjects believed that the partner was a human participant playing the game in a neighbouring room (Figure 3.2). In actual fact, and in part motivated by a need for good experimental control, I delivered feedback of a simulated player with preferences very different from the subjects' own.



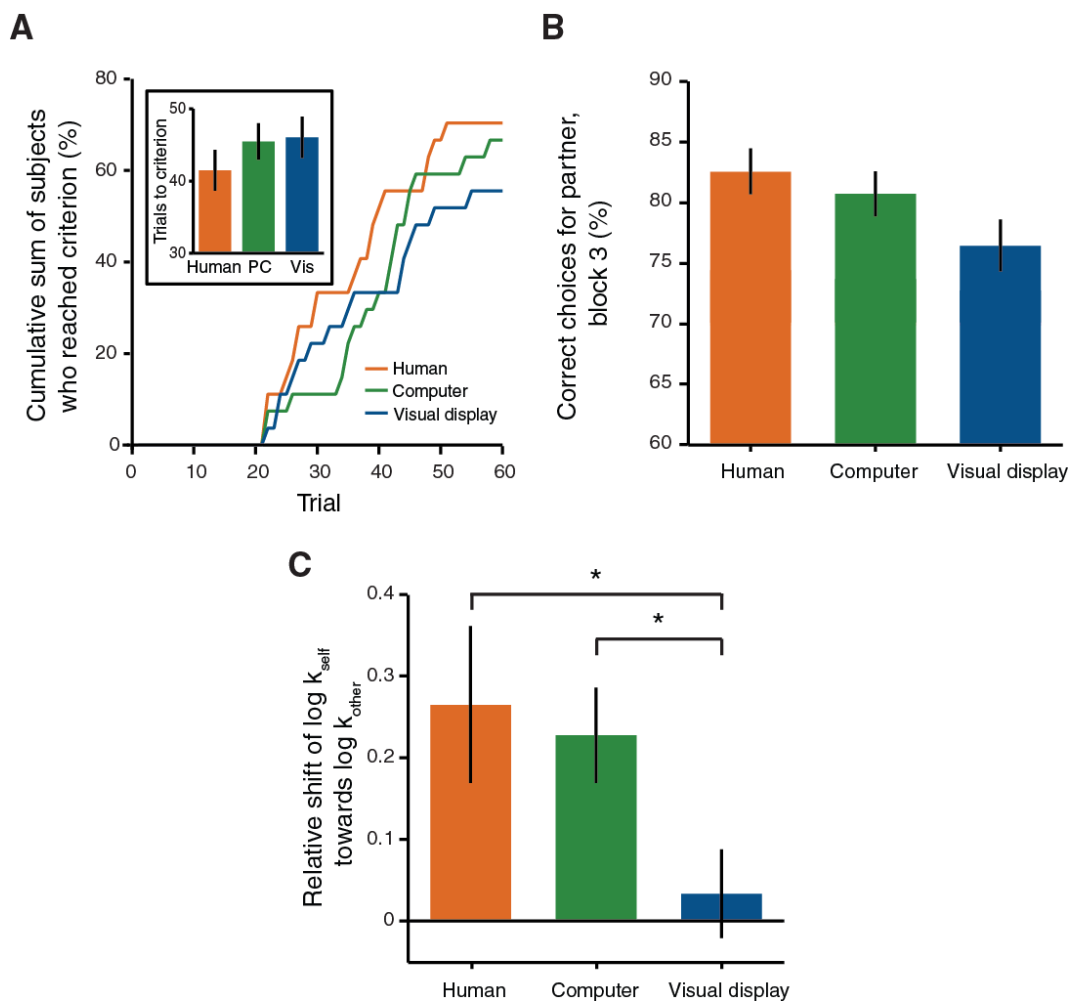


**Figure 3.8 Estimation of belief about the other's discount rate in one example subject.** **A** White dot indicates parameter combination of the prior belief about the other's preferences, estimated from subjects' choices on behalf of the other. White dashed line indicates true discount rate of the other. Black dashed line indicates this subject's own discount rate in block 1. **B** Evolution of estimated discount rate (orange) and prediction error (green) over trials. Black dashed line indicates true discount rate of the other. **C** Marginal posterior distribution tracking the belief about the other's discount rate across trials. The dashed white line indicates the true discount rate of the other. **B** and **C** demonstrate that the Bayesian learner was confident that the other's discount rate was stable at around  $\log k = -2$  after approximately 30 trials.

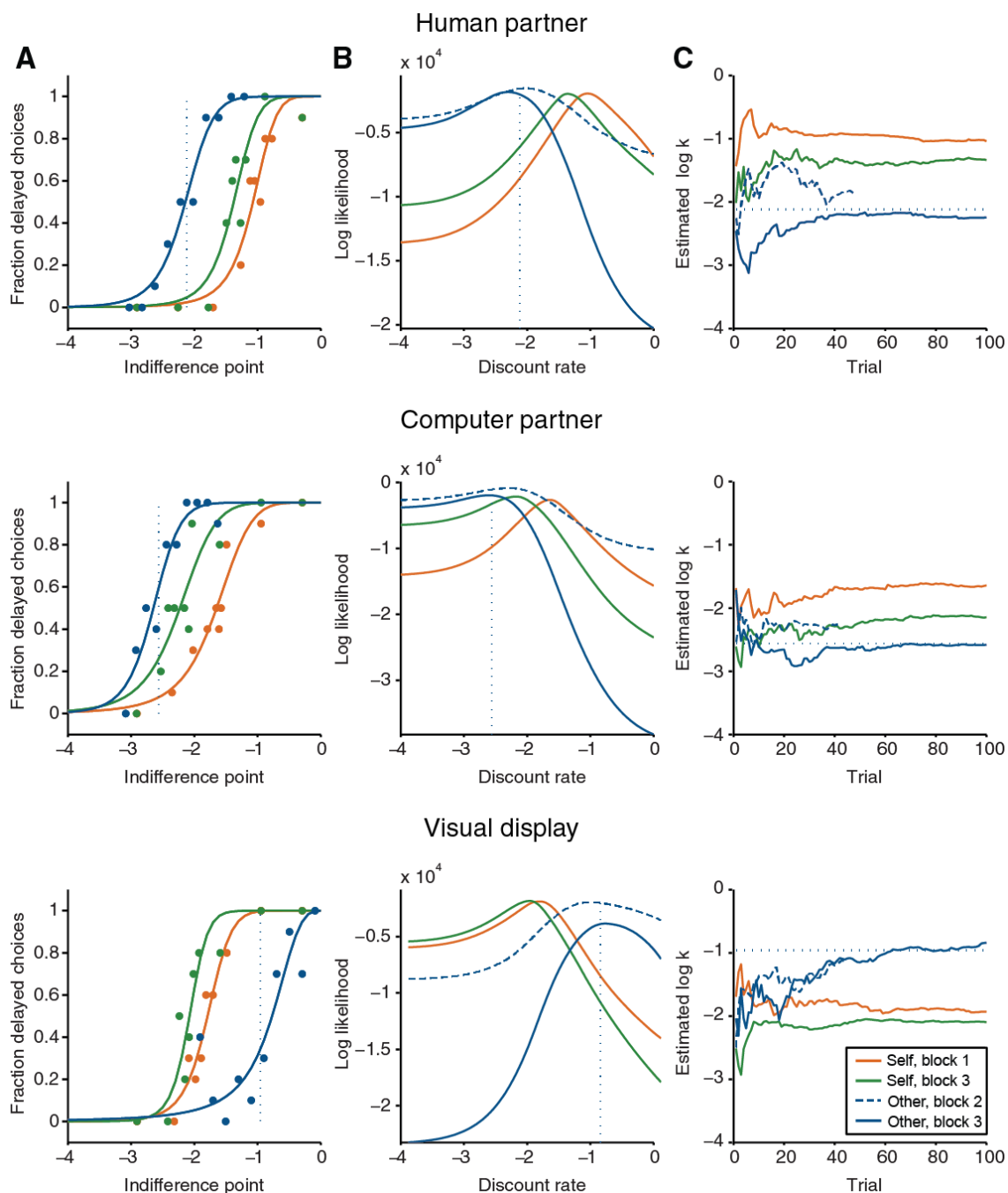
The discount rate used to generate the other's choices could not be inferred from individual trials, and choices of the confederate were noisy. Optimal behaviour in this task therefore required subjects to learn the preferences of the other individual by tracking their behaviour across trials and updating beliefs when a new piece of information was received at the feedback stage. The integration of knowledge across trials can be modelled using a Bayesian learner whose initial prior distribution  $\log N(k_0, \sigma_0^2)$  was first estimated (Figure 3.8A) and then updated on every trial (Figure 3.8B,C). Ultimately, the estimate of the other's

discount rate approximated the actual discount rate used to generate choice behaviour for the other. Belief updates are likely to be driven by prediction errors, arising at the feedback stage when a subject's choice for the other is inconsistent with the other's actual decision. Indeed, changes in belief are particularly large on trials where a prediction error is encountered (Figure 3.8B). The size of the prediction error here is quantified as the discrepancy between a subject's estimated discount rate before feedback and the estimated discount rate after feedback (Figure 3.8B).

### 3.4.2 Discount rates are susceptible to social influence

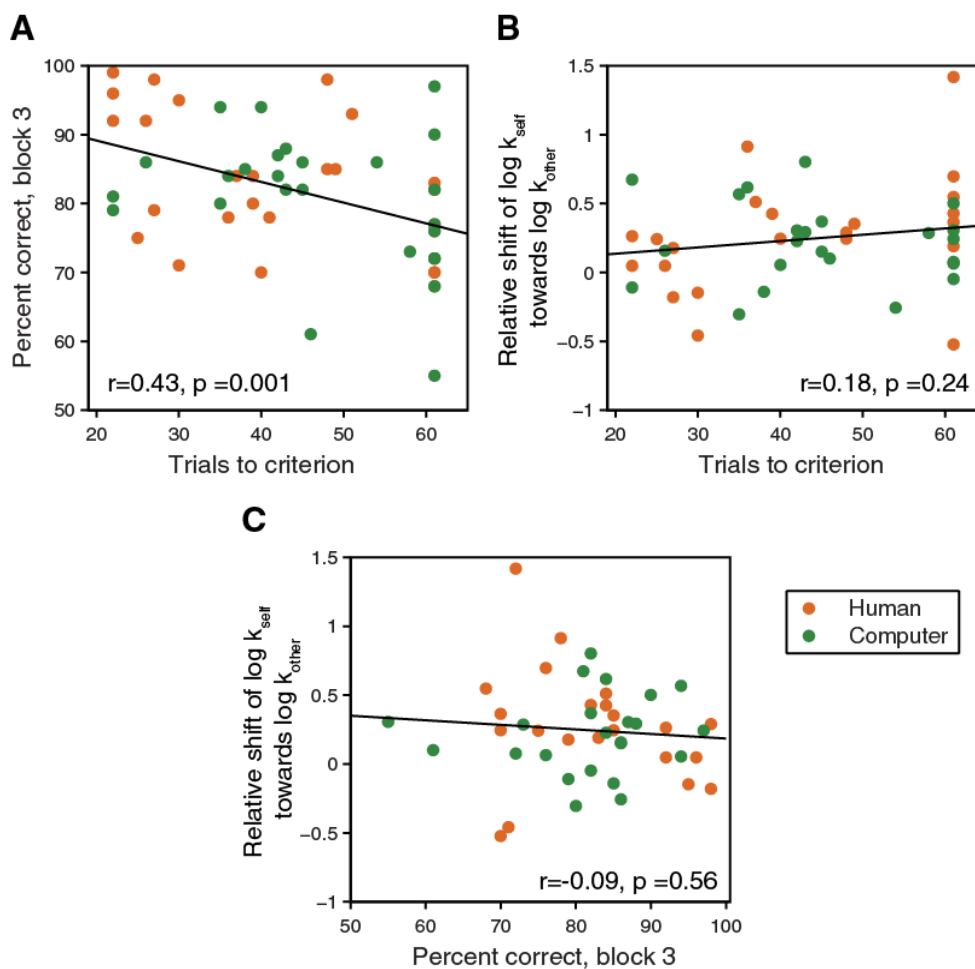


**Figure 3.9 Learning the discount rate of another and the consequences for behaviour.** **A** Block 2 terminated after 17 correct choices for the confederate within a sliding window of 20 consecutive trials or a maximum of 60 trials. Depicted here is the cumulative sum of participants who terminated after a given number of trials. The inset shows the mean number of trials to criterion for the different groups. The average number of trials subjects needed to reach criterion in block 2 did not differ between groups ( $F_{2,78} = 0.82$ ,  $P = 0.44$ ). **B** Correct choices for the confederate in block 3 did not differ between groups ( $F_{2,78} = 2.55$ ,  $P = 0.08$ ), suggesting that subjects in all three groups learnt the other's discount rate equally well. **C** Shift of subjects' own discount rate in the direction of the partner's discount rate relative to the distance between  $= \log k_{\text{self,block 1}}$  and  $\log k_{\text{other,block 2}}$ .



**Figure 3.10** Visualization of choice behaviour in one representative subject in the human partner group (top), the computer partner group (middle) and the visual display group (bottom). **A** Fraction of delayed choices as a function of the indifference point a choice can be mapped onto. For example, the decision [£10 today OR £100 in 3 months] maps onto -1, as an individual with a discount rate of -1 would be indifferent in this situation. [£10 today OR £11 in 3 months] maps onto -3 according to the same logic. **B** Summed likelihood distribution across discount rates. The peak indicates the estimate of the subject’s own or the other’s discount rate. **C** Evolution of discount rate estimate over trials. In all subplots, [self, block 1] is depicted in orange, [self, block 3] is depicted in green, [other, block 2] is depicted in blue (dashed line) and [other, block 3] is depicted in blue (solid line). Dashed blue line indicates the modelled other’s discount rate.

As previously reported (Nicolle et al., 2012), subjects learnt quickly, and accurately, to choose according to a novel partner's preferences (Figure 3.9A,B). Notably, found that, after learning a partner's preferences, subjects' own discount rate shifted in the direction of the partner (relative shift calculated as  $\frac{\log k_{\text{self,block 3}} - \log k_{\text{self,block 1}}}{\log k_{\text{other,block 2}} - \log k_{\text{self,block 1}}}$ ,  $t_{21} = 3.06$ ,  $P = 0.006$ , visualized for individual subjects in Figure 3.10 and for the entire group in Figure 3.9). Their estimate of the novel other's preferences remained stationary (relative shift of other's discount rate calculated as  $\frac{\log k_{\text{other,block 3}} - \log k_{\text{other,block 2}}}{\log k_{\text{self,block 1}} - \log k_{\text{other,block 2}}}$ ,  $t_{21} = 0.99$ ,  $P = 0.33$ ) and was not biased towards subjects' own preferences ( $t_{21} = 0.49$ ,  $P = 0.63$ ).



**Figure 3.11 Relationship between behavioural shift in preference and performance on ‘other’ trials.** **A** Correlation between the number of ‘other’ trials to criterion in block 2 and performance for ‘other’ in block 3. **B** Correlation between number of ‘other’ trials to criterion in block 2 and the shift of subjects’ own discount rate in the direction of the partner’s discount rate relative to the distance between  $\log k_{\text{self,block 1}}$  and  $\log k_{\text{other,block 2}}$ . **C** Correlation between performance for ‘other’ in block 3 and shift of  $\log k_{\text{self}}$  towards  $\log k_{\text{other}}$ .

This effect is not easily understood as a social norm effect (Ruff et al., 2013) as discount rates shifted similarly when subjects were instructed they were deciding on behalf of a

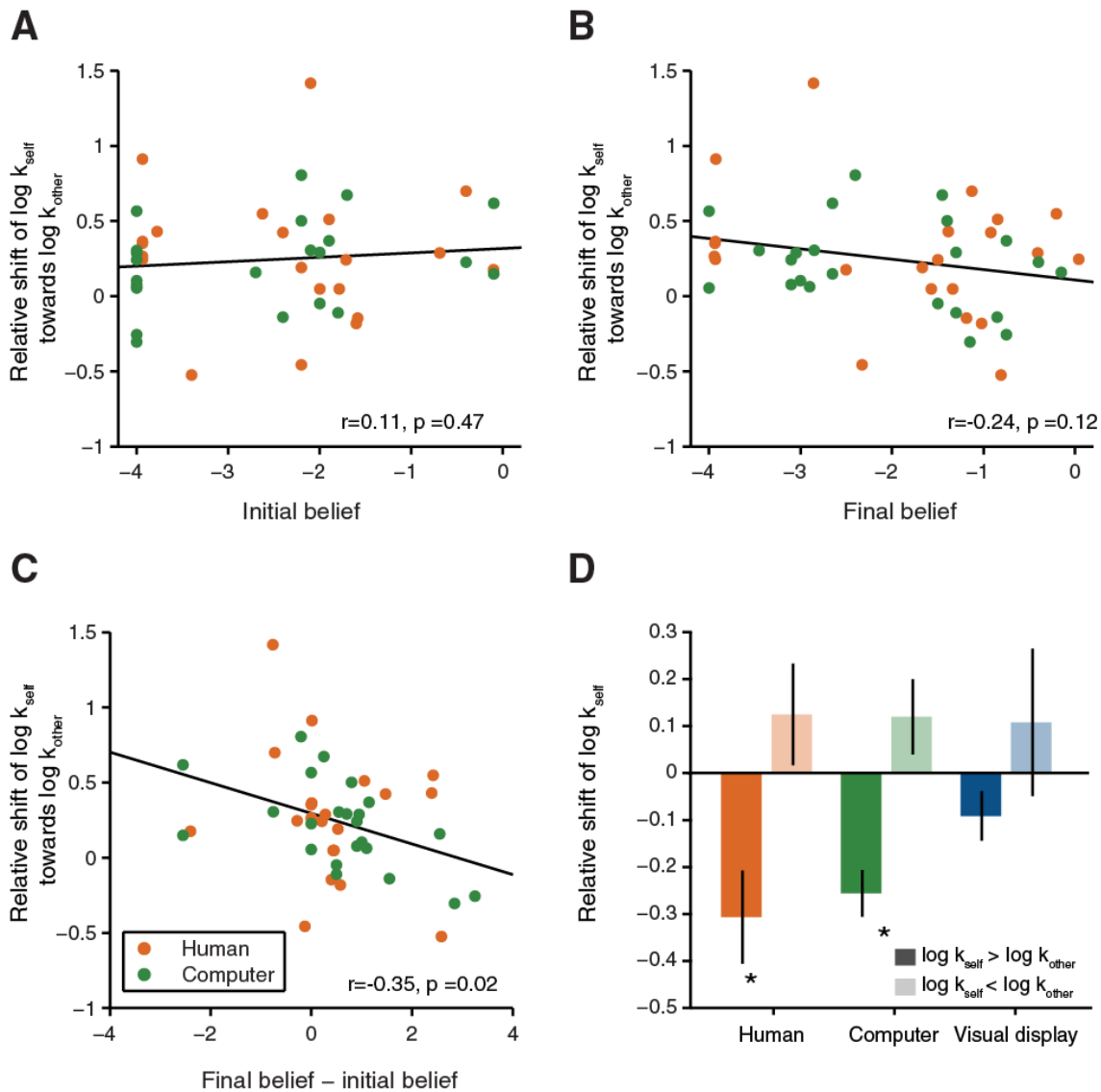
computer agent ( $t_{22} = 3.89$ ,  $P < 0.001$ , Figure 3.9C, Figure 3.10 middle). Subjects' estimate of the novel other's discount rate did not change from block 2 to block 3 in either of the two conditions (relative shift of other's discount rate calculated as  $\frac{\log k_{\text{other,block 3}} - \log k_{\text{other,block 2}}}{\log k_{\text{self,block 1}} - \log k_{\text{other,block 2}}}$ , human partner:  $t_{21} = 0.99$ ,  $P = 0.33$ , computer partner:  $t_{23} = -1.75$ ,  $P = 0.09$ ). This confirms that discount rates for self and other are not converging. Instead the change in discount rate corresponds to a selective shift of subjects' own discount rate towards the discount rate of the other.

Although the number of trials to criterion as well as the performance for the other in block 3 varies across subjects, this variability is not predictive of an individual's shift in preference (Figure 3.11). Similarly, the preference shift is not related to subjects' initial belief about the other's discount rate (Figure 3.12A), or the final belief at the end of the learning period (Figure 3.12B). However, those subjects whose beliefs change negatively (i.e. the other is more patient than they initially believed) shift more towards the novel other than those subjects whose beliefs change positively (i.e. the other is more impatient than they initially believed (Figure 3.12C). This is in line with the observation that subjects' shift was also particularly pronounced if the other was more patient than they were themselves (Figure 3.12D).

### 3.4.3 Discount rate shifts arise out of a simulation of the other's preferences

One account of this shift in preference is that it arises out of a simulation of the other's preferences. In order to test whether such simulation is crucial for this shift or whether the behaviour can be explained by simple stimulus- or action-based reinforcement, we designed a category-learning control experiment (Ashby and Maddox, 2005). This consisted of the same stimuli and actions, but the necessity to simulate another's discount rate was removed. Subjects were presented with a geometric depiction of a given choice on the screen (x-axis: delay of the later option, y-axis: ratio of magnitudes  $M_{LL}/M_{SS}$ , Figure 3.5, right) and instructed to choose according to the location of the dot with respect to an imaginary isoproprobability line. Rather than using feedback to update a value simulation, subjects now updated their belief about the orientation of this line. There was no differences in the number of trials required to reach criterion ( $F_{2,78} = 0.82$ ,  $P = 0.44$ ) or performance on 'other' trials in block 3 ( $F_{2,78} = 2.55$ ,  $P = 0.08$ ) compared to the human and computer partner conditions suggesting that learning and performance were comparable. However, in this scenario, subjects'

discount rates did not shift, indicating that subjects were not merely repeating previous choices they had made on behalf of the other ( $t_{24} = 0.61$ ,  $P = 0.55$ , Figure 3.9).



**Figure 3.12 Relationship between belief about the other's discount rate, and subjects' shift in preference.** **A** Correlation between subjects' initial belief about the other's discount rate and the shift of their own preferences towards the preferences of the novel other. **B** Correlation between subjects' final belief about the other's discount rate at the end of block 2 and the shift of their own preferences towards the preferences of the novel other. **C** Correlation between the difference in subjects' final belief and subjects' initial belief about the other's discount rate and the shift of their own preferences towards the preferences of the novel other. **D** Shift of  $\log k$  separately for situations where the other had a smaller, or a larger discount rate than the self.

This was confirmed in a one-way ANOVA, which revealed that the shift towards the other's preferences differed between experimental groups ( $F_{2,68} = 3.5$ ,  $P = 0.04$ ). Post-hoc t-tests attributed this difference to a smaller shift in the visual display group: Both subjects in the human and in the computer partner group displayed a stronger shift in discount rate

towards the other than subjects in the visual display group ( $t_{45} = 2.37$ ,  $P = 0.02$  and  $t_{47} = 2.25$ ,  $P = 0.03$ ). There was no difference in shift towards the partner for the human versus the computer partner group ( $t_{44} = 0.61$ ,  $P = 0.5$ ). Figure 3.10 visualizes the results for three representative individuals.

Note we cannot rule out differences in terms of attention, working memory or other factors that prevent the update of one's own preferences for the visual display condition. Since stimuli and actions were the same as in the human and the computer partner condition, however, we can rule out that the behavioural shift in the other experimental conditions is due to simple stimulus- or action-reinforcement. Instead, preference simulation is necessary to induce modulation in a discount rate.

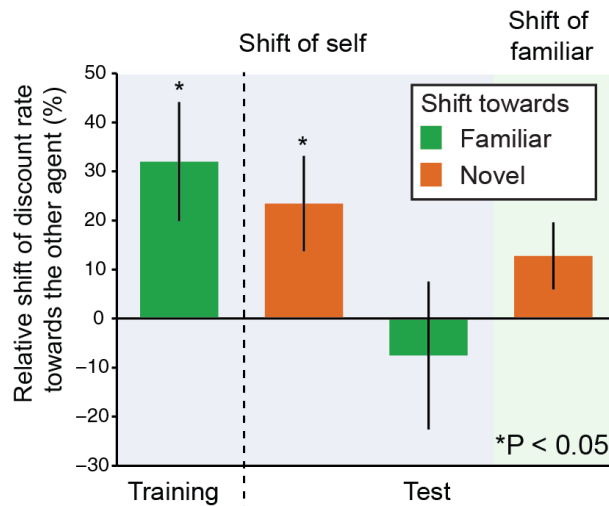
#### **3.4.4 Subjective value changes are induced by learning**

To be certain that any effects were driven by learning about the partner, as opposed to exercising a choice per se, we designed another version of the behavioural experiment, in which the discount rates of two partners were learnt at different points in time. This experiment comprised three players: the subject ("self"), a partner whose preferences were learnt in a training session ("familiar other") and a partner whose preferences were learnt during the actual test phase ("novel other"). The familiar and novel others' choices were simulated based on discount rates placed on opposite, and counterbalanced, sides of the subject's original discount rate. This means that one partner had a smaller, and the other partner a larger, discount rate than the subject himself.

16 participants performed two interleaved tasks. In choice trials, as in the behavioural experiment described above, participants again made inter-temporal choices for themselves and for the two partners. In probe trials, participants performed evaluations serially on behalf of different players, which would allow us to measure repetition suppression between the value representations of different individuals in a later fMRI experiment (Chapter 4). After each choice trial for the novel or the familiar partner, but not after probe trials, subjects were given feedback about the choice the confederate had made.

In this context, subjects' discount rates shifted towards the discount rate of the familiar partner during the initial training ( $t_{14} = 2.63$ ,  $P = 0.02$ ). During test, subjects' own discount rate shifted towards the newly learnt discount rate of the novel partner ( $t_{12} = 2.41$ ,  $P = 0.03$ ), but not the discount rate of the familiar partner ( $t_{11} = 0.50$ ,  $P = 0.63$ ). Furthermore, the shift of subjects' estimated discount rate of the familiar partner, shifted slightly towards the newly

learnt discount rate of the novel partner, but this shift did not reach significance ( $t_{13} = 1.86$ ,  $P = 0.08$ ). These preference shifts were therefore not simply associated with repeating the partner's choices but instead are most parsimoniously explained as induced by learning a new individual's preferences.



**Figure 3.13 The shift in preference is induced by learning about the other's preferences.** In the 2-partner version of the experiment, participants are first trained on the preferences of one partner (familiar other). This training induced shifts of the participants' own discount rates in the direction of the interaction partner's discount rate (green part/training,  $t_{14} = 2.63$ ,  $P = 0.02$ ). During the main part of the experiment, participants' discount rate moved towards the discount rate of the partner they now learned about (novel other,  $t_{12} = 2.41$ ,  $P = 0.03$ ), but not towards the familiar other ( $t_{11} = 0.50$ ,  $P = 0.63$ ). Subjects' estimate of the familiar other's discount rate shifts towards the novel did not reach significance ( $t_{13} = 1.86$ ,  $P = 0.08$ ).

### 3.5 Discussion

Here we investigated the behavioural consequences of learning about another's preferences in a delegated intertemporal choice task. In line with previous reports that highlight a social influence on the valuation of objects (Campbell-Meiklejohn et al., 2010; Klucharev et al., 2009; Zaki et al., 2011), we found that participants' own discount rate shifted towards the discount rate of a human or computer partner when learning about another's temporal discounting preferences. Importantly, this was only the case if we encouraged the simulation of the partner's preferences, but not if the same decisions were made based on a geometric display of the choice options. This demonstrates that participants integrate information about intertemporal preferences into their own preferences only if they use the same psychological mechanism to make choices for the partner as they do for themselves. On this basis we conclude that the observed change in subjective preference is best



understood as a consequence of a learning-related plasticity in neural populations computing valuations for self. Such a phenomenon could arise as a consequence of an overlap in the population of neurons in mPFC computing valuation for self and for others (Jenkins et al., 2008; Nicolle et al., 2012; Suzuki et al., 2012). The plasticity caused by learning about the preferences of a novel other may concurrently affect other value computations, such as subjects' own.

Subjects' own discount rate shifted towards the discount rate of their partner, irrespective of whether their partner was human or a computer. This is in line with studies demonstrating that individuals use strategies akin to those used in real social contexts when interacting with a computer agent (Nass and Moon, 2000). Crucially, a control condition with the same stimuli and actions, but without the need to employ a discounting computation, did not evoke a change in subjects' own preferences. This demonstrates that the shift cannot be explained by social rewards or expected social approval for making choices like another person. Instead, the behavioural effect is tied to subjects' deployment of the very same discounting mechanism to learn on behalf of another agent, be it a human or non-human agent. Furthermore, when subjects made decisions on behalf of two agents – a familiar other whose preferences they had learnt about previously, and a novel other whose preferences they currently learn about – their own discount rate shifted specifically towards the discount rate of the novel other. Thus, it is presumably a learning-induced plasticity in acquiring a novel value representation that impacted on subjects' own subjective value computation rather than the execution of a choice for another person alone.

The discount rate parameter manipulated here describes the degree to which future rewards are devalued (Myerson and Green, 1995) and is a reliable measure of impulsivity (Evenden, 1999; Robbins et al., 2012). It is known to be elevated in drug addiction, problem gambling, and other impulsivity disorders (Madden et al., 1997). In the absence of experimental manipulations, an individual's discount rate is often considered to be stable over time (Kirby, 2009; Ohmura et al., 2006). However, our results hint that in the real world social context plays an important role in determining impulsivity. The existence of such an influence raises the possibility of self-reinforcing patterns of impulsivity within social groups. Indeed, the magnetic effect exerted on a participant's discount rate by another person provides an interesting mechanistic link to the well-known phenomenon of social context influencing relapse, or alternatively self-restraint, in substance abuse.

Social conformity effects are also prominent in a range of other contexts (Campbell-Meiklejohn et al., 2010; Edelson et al., 2011; Zaki et al., 2011) and they may serve different cultural functions. For one, it has repeatedly been argued that we owe the cultural success of mankind to our exceptional ability to learn from others (Boyd et al., 2011; Frith and Frith, 2010). There may be good reason to believe that the other has more information about the world than me, in which case adapting my preferences to align with his would be adaptive. Crucially, this would mean that a preference shift should be less pronounced in a context where the other's preferences are not adaptive. Secondly, aligning preferences can be very important for collective decision making. The ability to work together towards common goals regularly requires reaching a consensus between individuals. This holds both for big political questions (Is it more important for the government to invest in health or in education?) as well as for less consequential personal decisions (Do we want to have sushi or pizza for dinner?). Decision making in such situations is greatly facilitated if the involved parties share values and preferences. A mechanism whereby humans approximate each other's values in a social setting could therefore be essential for decision making in groups.

# **4 LEARNING-INDUCED PLASTICITY IN MEDIAL PREFRONTAL CORTEX PREDICTS PREFERENCE MALLEABILITY**

\* This chapter is published in the following article:

Garvert MM, Moutoussis M, Kurth-Nelson Z, Behrens TEJ & Dolan RJ (2015).  
Learning-induced plasticity in medial prefrontal cortex predicts preference  
malleability. *Neuron* 85: 418-428

## 4.1 Abstract

Learning induces plasticity in neuronal networks. As neuronal populations contribute to multiple representations, we reasoned plasticity in one representation might influence others. We used human fMRI repetition suppression to show that plasticity induced by learning another individual's values impacts upon a value representation for oneself in medial prefrontal cortex (mPFC), a plasticity also evident behaviourally in a preference shift. We show this plasticity is driven by a striatal "prediction error", signalling the discrepancy between the other's choice and a subject's own preferences. Thus, our data highlight that mPFC encodes agent-independent representations of subjective value, such that prediction errors simultaneously update multiple agents' value representations. As the resulting change in representational similarity predicts inter-individual differences in the malleability of subjective preferences, our findings shed new mechanistic light on complex human processes such as the powerful influence of social interaction on beliefs and preferences.

## 4.2 Introduction

Information in the brain is encoded within distributed neuronal populations such that individual neurons typically support more than one representation or computation. Neurons in medial prefrontal cortex (mPFC), for example, perform self-referential as well as social value computations (Jenkins et al., 2008; Nicolle et al., 2012; Suzuki et al., 2012). Whereas it is traditionally suggested that computations for self and other are performed within separate populations of neurons (D'Argembeau et al., 2007; Denny et al., 2012), recent work suggests a functional organization within this region does not neatly conform to such a distinction by agent. Instead value computations on behalf of any individual can be realised by the same circuitry (Nicolle et al., 2012) and the neural code depends only on the subjective value of an offer. In light of this we conjectured that multiple value computations might be updated simultaneously if plasticity is introduced into this circuitry.

The contribution of overlapping neural circuitry to distinct computations has previously been demonstrated during delegated inter-temporal choice (Nicolle et al., 2012). In inter-temporal choice paradigms, subjects reveal their preferences for larger rewards delivered later versus smaller rewards that arrive sooner. Choice in this context is quantified by a 'temporal discount rate' (Myerson and Green, 1995), believed to index forms of behavioural impulsivity

(Evenden, 1999; Robbins et al., 2012) and an ability to imagine future outcomes (Cooper et al., 2013; Ersner-Hershfield et al., 2009; Mitchell et al., 2010; Peters and Büchel, 2010). When subjects are asked to make such inter-temporal choices on behalf of another individual (“delegated inter-temporal choice”), they rapidly learn the confederate’s discount rate (Nicolle et al., 2012). This adaptability depends on the medial frontal cortex, where a neural circuitry used to compute a subject’s own values also computes those of a confederate, enabling rapid switches between the two computations (Nicolle et al., 2012).

We reasoned that if the same circuitry in the mPFC computes the value of a delayed offer irrespective of agents, plastic changes necessary to learn a new partner’s preferences might have consequences for a subject’s own value computations. The presence of such plasticity would also be expected to induce behavioural change in the subject’s own temporal discount rate, a parameter usually assumed to index a stable personality trait (Kirby, 2009; Ohmura et al., 2006). One can conjecture that such plasticity might underlie social conformity effects, where individuals adjust their beliefs or preferences to align more with those with whom they interact (Campbell-Meiklejohn et al., 2010; Edelson et al., 2011; Zaki et al., 2011).

At a neuronal level a formal test of these predictions requires a fine-grained access to neural populations supporting distinct value computations, as well as a robust measure of learning-induced change in activity of these same populations. Despite its coarse spatial resolution, fMRI can reveal relationships between underlying cellular representations. In particular, fMRI adaptation paradigms can be finessed to measure plastic changes associated with the behavioural pairing of different items (Barron et al., 2013; Klein-Flügge et al., 2013b). The principle of fMRI adaptation builds on the idea that the repeated engagement of the same neuronal population leads to a diminished response and attenuated BOLD signal, even though the underlying biophysical mechanism remains ambiguous (Grill-Spector et al., 2006; Kohn, 2007).

Here we used an fMRI adaptation paradigm to measure the relationship between neuronal value representations for self, a familiar other whose preferences had been learnt prior to scanning and a novel confederate as this latter agent’s preferences were learnt. We deployed a dynamic repetition suppression procedure to provide us with a probe of plastic neural changes associated with learning a new flexible computation. We hypothesized that plasticity associated with this new learning would impact upon the preference representation for self as a consequence of a neuronal representation that maps agent and offer onto an agent-independent measure of subjective value. In essence this predicts that neuronal value

representations between self and a novel other should become more similar with learning, in line with a behavioural shift in preference. An alternative hypothesis posits separate value computations for distinct agents. In such a case a subject might use their own separate neural representations as a proxy for understanding another's traits, and an independent neuronal value representation for this other would be constructed through learning-induced plasticity (Barron et al., 2013). This alternative scenario predicts that neural value representations for self and other should become less similar with learning. In terms of a mechanism driving such plasticity, we reasoned that the same prediction errors that drive learning about a new partner's inter-temporal preferences would also induce shifts in the subject's own discount rate towards that of the partner.

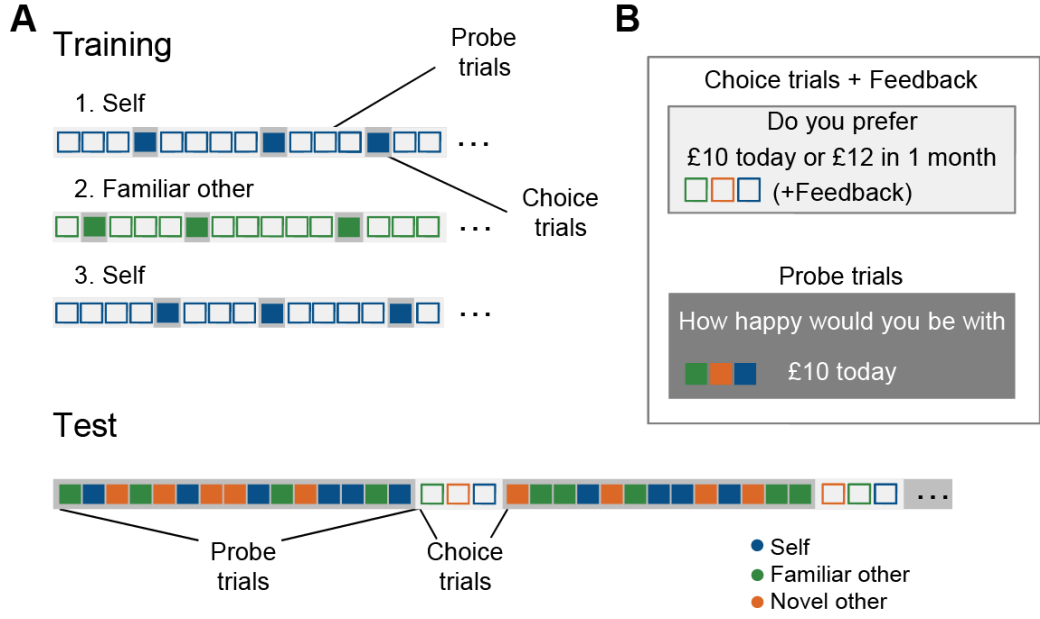
## 4.3 Methods

### 4.3.1 Subjects

29 volunteers (mean age  $\pm$  std:  $25.6 \pm 5.6$  years, 14 females) participated in the fMRI experiment. Two subjects were excluded from fMRI analyses, one because they had previously participated in the behavioural experiment and a second subject because of technical difficulties during data acquisition. All subjects were neurologically and psychiatrically healthy. The study took place at the Wellcome Trust Centre for Neuroimaging in London, UK. The experimental procedure was approved by the University College London Hospitals Ethics Committee and written informed consent was obtained from all subjects.

### 4.3.2 Task– fMRI study

The fMRI experiment was identical to the 2-partner behavioural experiment described in Chapter 3 and consisted of two trial types: choice trials and probe trials, in which subjects evaluated a single option on a scale from 1 to 4 (Figure 4.1). Subjects learned the preferences of a second partner ('familiar other') before the scan (Figure 4.1A). Option generation and shift estimates were the same as described in Chapter 3.



**Figure 4.1 Experimental design of the fMRI experiment.** **A** During training, blocks 1 and 3 consisted of self choice and probe trials, and block 2 consisted of ‘familiar other’ choice and probe trials. During test, subjects chose and evaluated for themselves, for the familiar other and for a novel other. The experiment was divided into three experimental blocks with probe trials the predominant type in all blocks. **B** Trial types. On choice trials, subjects chose between an immediately available, smaller and a delayed, larger reward. On “self” trials, subjects considered the choice for themselves. On “other” trials, they made the choice on behalf of a partner, and feedback indicated whether their choice corresponded to the partner’s (simulated) choice. On probe trials subjects indicated on a four-item scale how happy they themselves or one of their partners would be with the presented option.

Again, we excluded the first third of all choice trials subjects performed in the scanner when estimating  $k_{self,scan}$ ,  $k_{novel,scan}$  and  $k_{familiar,scan}$ . The relative shift effects reported in were then calculated as:

$$\begin{aligned} shift_{self \rightarrow fam, train} &= \frac{\log k_{self, training\_block\ 3} - \log k_{self, training\_block\ 1}}{\log k_{familiar, training\_block\ 2} - \log k_{self, training\_block\ 1}} \end{aligned} \quad (4.1)$$

$$shift_{self \rightarrow fam, test} = \frac{\log k_{self, test} - \log k_{self, training\_block\ 3}}{\log k_{familiar, test} - \log k_{self, training\_block\ 3}} \quad (4.2)$$

$$shift_{self \rightarrow novel, test} = \frac{\log k_{self, test} - \log k_{self, training\_block\ 3}}{\log k_{novel, test} - \log k_{self, training\_block\ 3}} \quad (4.3)$$

$$\text{shift}_{\text{fam} \rightarrow \text{novel}, \text{test}} = \frac{\log k_{\text{familiar}, \text{test}} - \log k_{\text{familiar}, \text{training\_block2}}}{\log k_{\text{novel}, \text{test}} - \log k_{\text{familiar}, \text{training\_block2}}} \quad (4.4)$$

As in the behavioural experiment, the estimation of absolute shifts the denominator  $z$  was set to  $\text{sign}(z)$ . Outliers (outside the range  $\text{mean} \pm 3 \cdot \text{std}$ ) as well as subjects for whom the denominator was smaller than 0.3 (2 subjects for  $\text{shift}_{\text{self} \rightarrow \text{fam}, \text{scan}}$ , 3 subjects for  $\text{shift}_{\text{self} \rightarrow \text{novel}, \text{scan}}$ , 2 subjects for  $\text{shift}_{\text{fam} \rightarrow \text{novel}, \text{scan}}$ ) were excluded from the analyses.

### 4.3.3 Surprise measure

I used the trial-by-trial estimate of subjects' own discount rates to compute differences in subjective value for all choices that subjects observed their partner make ( $V_{\text{chosen\_by\_partner}} - V_{\text{unchosen\_by\_partner}}$ ), where  $V_{\text{chosen\_by\_partner}}$  and  $V_{\text{unchosen\_by\_partner}}$  were computed according to equation (4.5):

$$V = \frac{M}{1 + kD} \quad (4.5)$$

This difference in subjective value was transformed to a probability indicating how likely the subject would be to make the same choice himself using a softmax function:

$$P(\text{chosen}) = \frac{1}{1 + e^{-\beta(V_{\text{chosen}} - V_{\text{unchosen}})}} \quad (4.6)$$

Subject's inverse temperature parameter  $\beta$  was also estimated on a trial-by-trial basis from subjects' choices on the task. This measure gave us an estimate of how likely the subject would have been to make the same choice himself. We subtracted this likelihood from 1 to translate this to a surprise measure.

### 4.3.4 Scan procedure, fMRI data acquisition and pre-processing

Visual stimuli were projected onto a screen via a computer monitor. Subjects indicated their choice using an MRI-compatible button box. Stimuli were presented for a minimum duration of 3 to 5 seconds or until subjects indicated their decision. MRI data was acquired using a 32-channel head coil on a 3 Tesla Allegra scanner (Siemens, Erlangen, Germany). A special sequence was used to acquire T2\*-weighted EPIs to minimize susceptibility related artefacts in the ventral prefrontal cortex (Weiskopf et al. 2006): 43 transverse slices



(ascending order) of 2 mm thickness with 1-mm gap and in-plane resolution of 3x3mm, a TR of 3.01s and TE of 70ms were collected. Slices were tilted by 30° relative to the rostro-caudal axis and a local z-shim with a moment of -0.4mT/m was applied to the OFC region. The first five volumes of each block were discarded to allow for equilibration. A T1-weighted anatomical scan with 1x1x1mm resolution was acquired at the end of the session in order to spatially normalize the EPIs. In addition, a whole-brain fieldmap with dual echo-time images (TE1 = 10ms, TE2 = 14.76ms, resolution 3x3x3mm) was obtained to correct for geometric distortions induced in the EPIs at high field strength.

We used SPM8 for image pre-processing and data analysis (Wellcome Trust Centre for Neuroimaging, London UK). We corrected for signal bias, co-registered functional scans to the first volume in the sequence and corrected for distortions using the fieldmap. Data were spatially normalized to a standard EPI template and smoothed using an 8mm full-width at half maximum Gaussian kernel.

#### **4.3.5 Physiological noise**

To reduce the contribution of physiological noise to the BOLD signal (Hutton et al. 2011), the cardiac pulse was recorded using an MRI compatible pulse oximeter (Model 8600 F0, Nonin Medical Inc., Plymouth, MN, USA) and thorax movement was monitored using a custom-made pneumatic belt positioned around the abdomen. The pneumatic pressure was converted into an analogue voltage signal using a pressure transducer (Honeywell International Inc., Morristown, NJ, USA) before digitization.

Models for cardiac and respiratory phase and their aliased harmonics were based on RETROICOR (Glover et al., 2000); the model for respiratory volume changes was based on (Birn et al., 2006). Slice 15 was used as a reference slice for modelling fluctuations arising from cardiac phase because of its proximity to the OFC (Hutton et al., 2011). Sessions were modelled separately within the general linear model (GLM).

#### **4.3.6 fMRI data analysis**

Data were analysed with an event-related GLM. Probe trials were sorted into nine different conditions (self preceded by self (SS), novel preceded by self (SN), familiar preceded by self (SF), self preceded by novel (NS), novel preceded by novel (NN), familiar preceded by novel (NF), self preceded by familiar (FS), novel preceded by familiar (FN), familiar

preceded by familiar (FF)) with 20 trials per condition and block. Each regressor was accompanied by a parametric modulator reflecting subjective value from the respective agent's perspective. This value was calculated based on a trial-by-trial estimate of the subject's current belief about their partners' discount rate  $k$ . Furthermore, we defined one choice regressor per agent and block indexing the time at which subjects indicated their decision on choice trials and received feedback. Each was accompanied by a parametric regressor corresponding to the surprise subjects experienced as they observed the partner's choice. Button presses were included as a regressor of no interest. Because of the sensitivity of the BOLD signal in the OFC region to subject motion and physiological noise, we included six motion regressors obtained during realignment as well as ten regressors for cardiac phase, six for respiratory phase and one for respiratory volume extracted with an in-house developed Matlab toolbox as nuisance regressors (Hutton et al., 2011). Sessions were modelled separately within the GLM.

To detect areas showing adaptation to repeated agents as depicted in Figure 4.3, we used the contrast [(agent preceded by different agent) – (agent preceded by same agent)], i.e.  $([SN + SF + NS + NF + FS + FN] - 2 \times [SS + FF + NN])$ . To test for areas displaying greater increases in suppression between self and the novel other compared to between self and familiar other (Figure 4.4A), we defined the following contrast:  $([SN+NS]_{\text{block1}} - [SN+NS]_{\text{block3}}) - ([SF+FS]_{\text{block1}} - [SF+FS]_{\text{block3}})$ . To test for greater increases in suppression between self and novel other than between novel other and familiar other, we defined a contrast as follows:  $([SN+NS]_{\text{block1}} - [SN+NS]_{\text{block3}}) - ([NF+FN]_{\text{block1}} - [NF+FN]_{\text{block3}})$ .

The contrast images of all subjects from the first level were analysed as a second-level random effects analysis. Results are reported at a cluster-defining threshold of  $P < 0.01$  uncorrected combined with a family-wise-error (FWE) corrected significance level of  $P < 0.05$ .

We performed a jack-knife procedure from the mPFC ROI (Figure 4.4A) to extract parameter estimates from this region without biasing the selection. To this end, we extracted parameter estimates for each subject from an ROI defined according to all other subjects (threshold at  $P < 0.01$  uncorrected). This signal was used to perform all analyses depicted in Figure 4.4-Figure 4.7 and Supplementary Figure 4.1.

In ROI correlation analyses, we performed partial correlations to control for correlations between  $\text{shift}_{\text{self} \rightarrow \text{novel, scan}}$  and  $\text{shift}_{\text{fam} \rightarrow \text{novel, scan}}$ . This removes the shift of the familiar other towards the novel other from the subjects' own discount rate shifts and the neural plasticity

index  $[\text{SN-SF}]_{1-3}$  (Figure 4.5A) and the shift of self towards the novel other from the familiar other's shift towards the novel other and the neural plasticity index (Figure 4.5B). We also estimated a linear regression model on the same data with  $\text{shift}_{\text{self} \rightarrow \text{novel,scan}}$  and  $\text{shift}_{\text{fam} \rightarrow \text{novel,scan}}$  as independent variables and  $[\text{SN-SF}]_{1-3}$  as the dependent variable. The relationship between  $\text{shift}_{\text{self} \rightarrow \text{novel,scan}}$  and  $[\text{SN-SF}]_{1-3}$  was directly contrasted with the relationship between  $\text{shift}_{\text{fam} \rightarrow \text{novel,scan}}$  and  $[\text{SN-SF}]_{1-3}$ .

To test for the influence of surprise on mPFC plasticity, we defined a contrast assessing BOLD correlate of the surprise subjects experienced as they got feedback about the novel and the familiar partners' choices. This contrast revealed activity in ACC, in bilateral insula and dorsal striatum (Figure 4.6A; note that insula and striatal activity did not survive cluster-based FEW thresholding). Parameter estimates were then extracted from these regions and correlated with the plasticity measure  $[\text{SN-SF}]_{1-3}$  extracted from the mPFC ROI in Figure 4.6B. To identify the surprise experienced when learning about the novel other, parameter estimates were then extracted from these ROIs for the novel other's choices only. This surprise measure in the striatum was correlated with subjects' shift in discount rate and the plasticity measure  $[\text{SN-SF}]_{1-3}$  (Figure 4.6B) extracted from the mPFC ROI (Figure 4.6C).

To test the specificity of adaptation effects we analysed repetition suppression effects in visual regions. We defined an ROI from a contrast identifying a main effect to any visual event, averaged across all blocks and performed the same analyses as for the mPFC ROI (thresholded at  $P < 0.0001$  uncorrected; Supplementary Figure 4.2).

#### 4.3.7 Mediation analysis

We used the Mediation and Moderation Toolbox (Atlas et al., 2010; Wager et al., 2008) to perform a single-level mediation analysis. To test whether mPFC plasticity mediates the effect of striatal surprise on behavioural shift, we first extracted each individual's parameter estimate from the striatal ROI encoding surprise. The mediator corresponded to each subject's plasticity index  $[\text{SN-SF}]_{1-3}$  computed from parameter estimates extracted from the mPFC ROI. The outcome variable was defined as a subject's relative shift in discount rate towards the novel other. The relationship between striatal surprise and behavioural shift controlling for the mPFC effect is referred to as path "c". We also estimated the relationship between striatal surprise and mPFC plasticity (path "a") as well as between mPFC plasticity and behavioural shift (path "b"). This last path "b" is controlled for striatal surprise, such that paths "a" and "b" correspond to two separable processes contributing to the behavioural

effect. A mediation test (path “ab”) examines whether the mediator (mPFC plasticity) explains a significant amount of the covariance between striatal surprise and behavioural shift. We determined two-tailed uncorrected P values from the bootstrap confidence intervals for the path coefficients (Atlas et al., 2010).

## 4.4 Results

### 4.4.1 Subjective value changes are induced by learning

The results described in Chapter 3 suggest that learning to compute the preferences of another agent induces plastic changes in the neural architecture responsible for personal valuation. This in turn predicts the neural population engaged during the computation of self-valuation should change over the course of the experiment. This population should either become closer to that evoked during valuation for the partner if the representational structure of an offer depends solely on its subjective value irrespective of the individual. Alternatively, it should become less close if separate agent-specific representations exist and subjects construct an independent representation for the novel other as a consequence of learning.

To test for such change in similarity between neural representations for self and others we interleaved trials from the delegated inter-temporal choice task with ‘probe’ trials. These probe trials enabled us to measure repetition suppression between individuals (Figure 4.1). We reasoned that, if self and partner valuation mechanisms overlapped more after learning than before, in line with an increase in behavioural similarity, then this predicts greater repetition suppression at the end of the experiment than at the beginning. If, however, subjects constructed a representation of the novel other from their representation of self, then this predicts the very opposite, namely repetition suppression at the beginning of the experiment which disappears as subjects build a separate representation of the novel partner.

Like the 2-partner behavioural experiment (Chapter 3), our experiment comprised three players: the subject (“self”), a partner whose preferences were learnt prior to scanning (“familiar other”) and a partner whose preferences were learnt during scanning (“novel other”). The familiar and novel others’ choices were simulated based on discount rates placed equally far apart on opposite, and counterbalanced, sides of the subject’s original discount rate. This meant that one partner had a smaller, and the other partner a larger, discount rate than the subject himself. The familiar partner was introduced to ensure that any effects were

driven by learning about the partner, as opposed to exercising a choice per se. The familiar other also controlled for non-specific time-dependent signal changes not associated with learning of new preferences.

I scanned 27 subjects whilst they performed the two interleaved tasks. In choice trials, as in the behavioural experiment described in Chapter 3, subjects again made inter-temporal choices for themselves and for the two partners. In ‘probe trials’, subjects performed evaluations serially on behalf of different players, allowing us to measure repetition suppression between the value representations of different individuals (Figure 4.1B). After each choice trial for the novel or the familiar partner, but not after probe trials, subjects were given feedback about the choice the confederate had made.



**Figure 4.2 The shift in preference is induced by learning about the other's preferences.** Relative shift of subjects' own discount rate (blue background) and the discount rate of the familiar other (green background) towards the familiar other (green bars) and the novel other (orange bars) during training and scanning. In line with the 2-partner behavioural experiment, training on the preferences of the familiar other induced shifts of the participants' own discount rates in the direction of the interaction partner's discount rate (green part/training,  $t_{23} = 3.17$ ,  $P = 0.004$ ). During scanning, participants' discount rate moved towards the discount rate of the partner they now learn about (novel other,  $t_{23} = 3.05$ ,  $P = 0.006$ ), but not towards the familiar other ( $t_{23} = -1.69$ ,  $P = 0.10$ ). Furthermore, subjects' estimate of the familiar other's discount rate also shifted towards the novel other ( $t_{24} = 2.87$ ,  $P = 0.008$ ).

In line with our previous behavioural results, in this fMRI based experiment subjects' discount rates shifted towards the discount rate of a familiar partner during preference

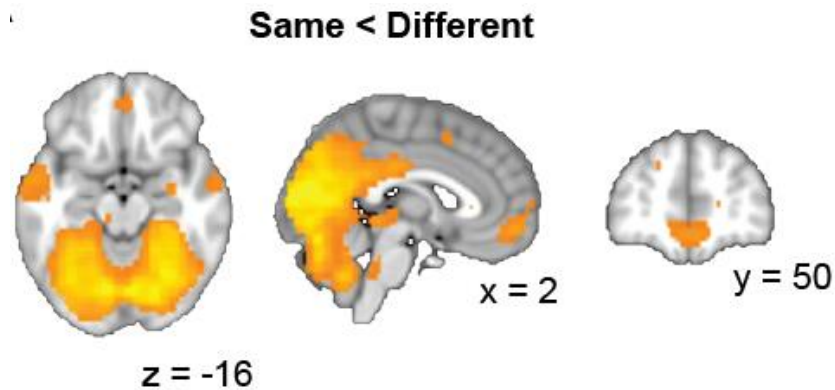
learning prior to scanning ( $t_{23} = 3.17$ ,  $P = 0.004$ ). During scanning, both subjects' own discount rate ( $t_{23} = 3.05$ ,  $P = 0.006$ ), and subjects' estimated discount rate of the familiar partner ( $t_{24} = 2.87$ ,  $P = 0.008$ ), shifted towards the newly learnt discount rate of the novel partner, with a stronger relative shift evident for subjects' own discount rate ( $t_{22} = 2.18$ ,  $P = 0.04$ , Figure 4.2), but comparable absolute shifts ( $t_{22} = 0.72$ ,  $P = 0.48$ ). These preference shifts were therefore not simply associated with repeating the partner's choices but instead are most parsimoniously explained as induced by learning a new individual's preferences.

#### 4.4.2 Plasticity between neural representations of self and other

To address whether a measured change in subjective preference is linked to plasticity in neural populations computing valuations for self, we focused our analysis on probe trials. We first established that we could measure repetition suppression by comparing brain activity elicited by simulating values for an agent when preceded by the same agent compared to a situation where an agent was preceded by another agent (Figure 4.3). Different agents were indicated to the subject by different colours on screen (Figure 4.1B). Unsurprisingly, we observed fMRI adaptation in the visual cortex ( $P < 0.001$ , peak  $t_{26} = 16.93$ ,  $[30, -61, -8]$ ), reported here and in subsequent fMRI analyses as family wise error (FWE) corrected on cluster level, (Figure 4.3, (Buckner et al., 1998; Wiggs and Martin, 1998), but also in a network that included mPFC ( $P = 0.02$ , peak  $t_{26} = 5.76$ ,  $[3, 53, -11]$ ) and left superior temporal sulcus (STS,  $P < 0.001$ , peak  $t_{26} = 4.95$ ,  $[-51, -13, -8]$ ). The latter two regions are associated with mentalizing (Gallagher and Frith, 2003), valuation for self (Boorman et al., 2009; Hunt et al., 2012; Kable and Glimcher, 2007) and valuation for others (Jenkins et al., 2008; Nicolle et al., 2012). Whilst this main effect of repetition suppression does not dissociate visual from agent-specific effects, it confirms that similarity in neural patterns evoked in a valuation network can be indexed by repetition suppression (Barron et al., 2013; Jenkins et al., 2008).

We reasoned that we could use this index of neural similarity to investigate whether the observed shift in subjective preferences was linked to plastic changes in the valuation network. If the neural code depends on the subjective values of a given offer alone, then repetition suppression should emerge between self and novel other over the course of the experiment, given that discount rates for self align with discount rates for a novel other. If, on the other hand, the mPFC encodes value differentially depending on agent, where learning another's preferences involves construction of an independent representation of this novel other from a representation of self, then repetition suppression should decrease over

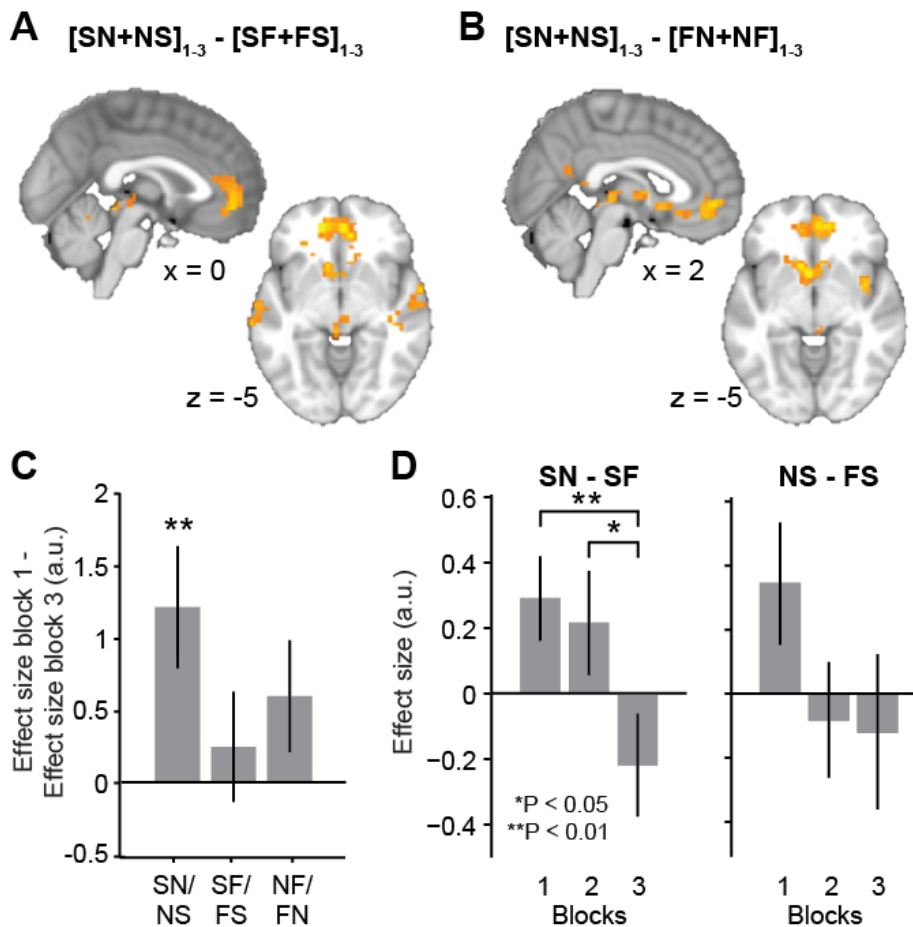
the course of the experiment. Whilst a similar change in suppression might also be predicted between novel and familiar others, there should be no such change in suppression between self and familiar other if in fact we are indexing changes induced by new learning.



**Figure 4.3 Repetition suppression as an index of representational similarity.** Displayed are brain areas with significantly less activity for repeated compared to changing agents on subsequent trials. Contrast images are thresholded at  $P < 0.01$  uncorrected for visualization.

We designed a contrast that measured the change in repetition suppression between self and novel other from block 1 to block 3, controlled for by the change in repetition suppression between self and familiar other over the same blocks. The only brain region to survive whole-brain statistical correction was in mPFC (Figure 4.4A,  $P = 0.01$ , peak  $t_{26} = 3.82$ ,  $[-12, 53, 1]$ ), although sub-threshold clusters in the left and right STS were also present ( $P = 0.27$ , peak  $t_{26} = 3.77$  and  $P = 0.48$ , peak  $t_{26} = 3.38$ , respectively). This region overlaps with an area involved in self-referential processing and in encoding value on probe trials (Supplementary Figure 4.1B,C). There were no significant effects for the opposite interaction. This change cannot be due to visual effects as we controlled for these both by inclusion of the familiar agent, and separately by the comparison between early and late blocks in the experiment. Consequently visual regions do not show these condition-specific changes in suppression (Supplementary Figure 4.2). Neither can the effect be due to novelty or differences in processing speed, as no differences between main effects of novel and familiar others were seen in this region (Supplementary Figure 4.1) or in the response times (Supplementary Figure 4.3). Furthermore, an equivalent contrast measuring the change in suppression between self and novel other, but now controlling for the change in suppression between familiar and novel other, revealed a similar change in activity in an overlapping brain region (Figure 4.4B). Hence, within the mPFC learning the preferences of a novel agent

specifically increased repetition suppression between representations of self and this novel partner.



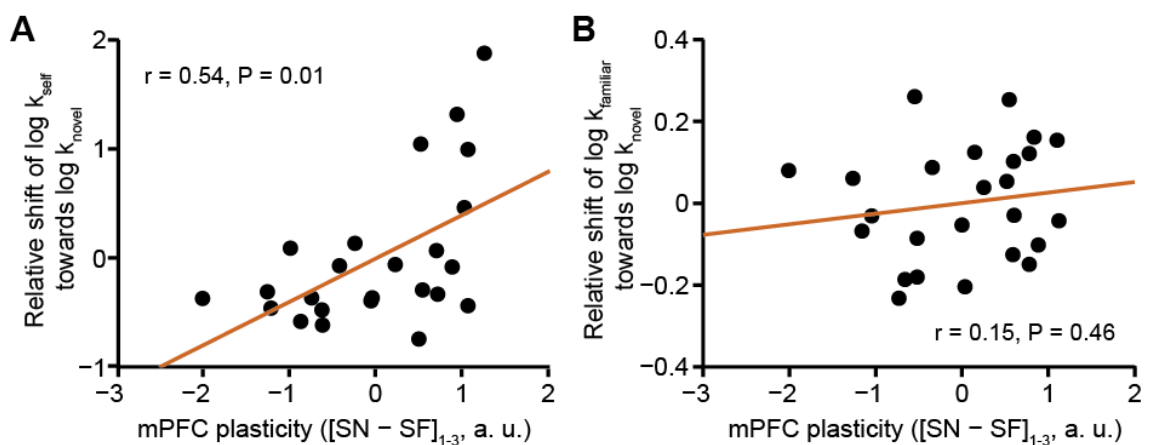
**Figure 4.4 Learning-induced plasticity in mPFC.** **A** Brain areas with a significantly greater increase in suppression from block 1 to block 3 between self and novel other compared to the increase in suppression between self and familiar other. **B** Areas displaying an increase in suppression from block 1 to block 3 between self and novel other relative to changes in suppression between novel and familiar other. **C** Parameter estimates extracted by a jack-knife procedure from the mPFC ROI in **A**, averaged across subjects. **D** Same parameter estimates as in **C** but now separated into the distinct components. Data are represented as mean  $\pm$  SEM. Contrast images in **A,B** are thresholded at  $P < 0.01$  uncorrected for visualization. a.u. arbitrary units

To further investigate mPFC suppression effects, we employed a jack-knife procedure across subjects to extract parameter estimates from the cluster of interest. Consistent with the whole-brain analysis, we found a significant change in novel-to-self/self-to-novel suppression (Figure 4.4C,  $t_{26} = 2.86$ ,  $P = 0.008$ ), but not in self-to-familiar/familiar-to-self suppression from block 1 to block 3 ( $t_{26} = 0.64$ ,  $P = 0.52$ ). The change in novel-to-familiar/familiar-to-novel suppression in the same ROI was in the same direction, but did not reach significance ( $t_{26} = 1.54$ ,  $P = 0.14$ ), and was smaller than the change in novel-to-



self/self-to-novel ( $t_{26} = 1.65$ ,  $P = 0.05$ ). Since overall activity in mPFC for self trials was greater than activity for other trials (Supplementary Figure 4.1), sensitivity to repetition suppression may differ depending on the order of the two agents. To explore potential differences, we decomposed the contrasts described above. Changes in repetition suppression between self and novel other were observed in both directions (Figure 4.4D), but were only significant when self trials were the priming and not the test trials (Figure 4.4D left, ANOVA: left,  $F_{2,78} = 3.39$ ,  $P = 0.04$ , right  $F_{2,78} = 1.55$ ,  $P = 0.21$ ).

#### 4.4.3 Plasticity in mPFC predicts discount rate shifts



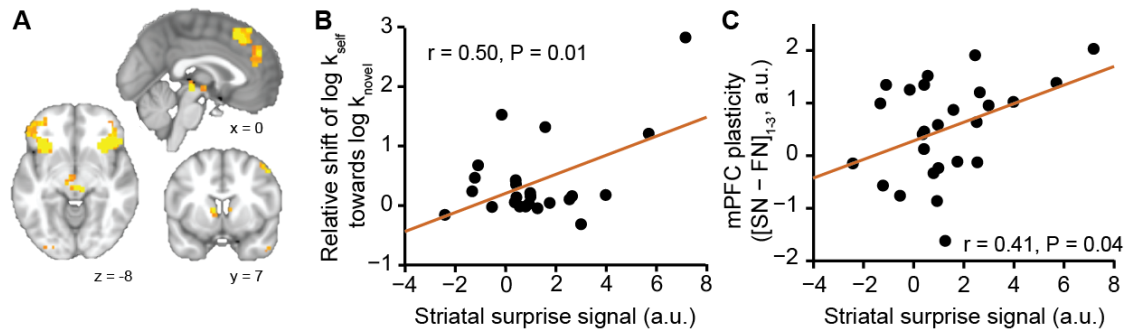
**Figure 4.5 Relationship between [SN-SF]<sub>1-3</sub> plasticity and shift in discount rate.** **A** Partial correlation between the magnitude of change in suppression between self and novel relative to the change in suppression between self and familiar agents over blocks and the shift in subjects' own discount rate towards the novel other. **B** Partial correlation between the change in suppression between self and novel relative to the change in suppression between self and familiar over blocks and the shift in subjects' estimate of the familiar other's discount rate towards the novel other. Parameter estimates in **A** and **B** were extracted from the mPFC ROI shown in Figure 4.4A. To account for the correlation between subjects' own shift in discount rate and the shift in their estimate of the familiar other's discount rate, we performed partial correlations, i.e. the familiar shift was removed from behaviour and neural signal in **A** and the self shift was removed from behaviour and neural signal in **B**. The relationship between [FN-SN]<sub>1-3</sub> plasticity and the shift of the familiar other's discount rate towards the novel other is analysed in Supplementary Figure 4.4. a.u., arbitrary units.

If the observed behavioural change in preference is related to learning-induced plasticity in value computations, then the increase in representational similarity between self and novel other should predict a subject's shift in preference. The increase in self-to-novel relative to self-to-familiar suppression over blocks did indeed predict the shift in subjects' own discount rate (partial correlation,  $r = 0.54$ ,  $P = 0.007$ , Figure 3A) towards the novel other, but not the same shift in the subjects' estimate of the familiar other's discount rate (partial correlation,  $r = 0.15$ ,  $P = 0.46$ , Figure 3B), although a direct comparison of these effects in a multiple

regression analysis did not reach significance ( $t_{23} = 0.71$ ,  $P = 0.24$ ). The shift in subjects' estimate of the familiar other's preferences was instead loosely related to an increase in representational similarity between familiar and novel other (Supplementary Figure 4.4). The fact that the behavioural estimate for a shift in discount rate was derived from choice trials, whereas the neural plasticity effect was extracted from probe trials strongly suggests that learning a partner's choice induces a stable plasticity in regions involved in value computation.

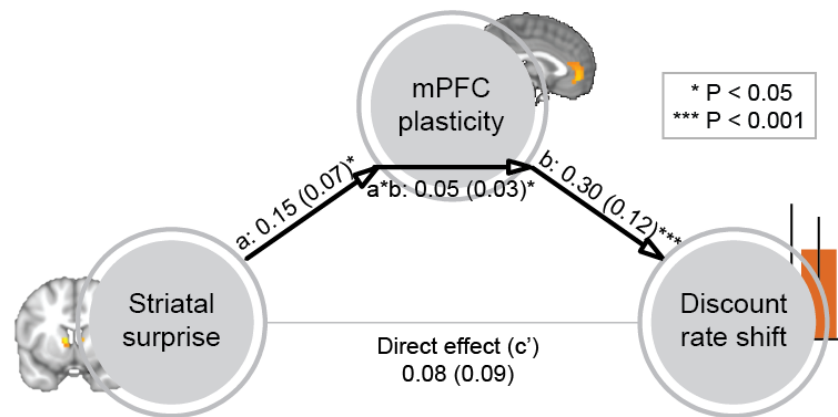
#### **4.4.4 Plasticity in mPFC is predicted by surprise coding in the striatum**

A plausible mechanism for inducing plastic change is surprise or prediction error, which in this context arises when the familiar or the novel partner's choices diverge from the choice the subject themselves would have made given the same choice context. Bayes-optimal estimates of this measure (see Experimental Procedures) were reflected in the posterior medial frontal cortex (Figure 4.6A,  $P = 0.04$ , peak  $t_{26} = 4.09$ ,  $[-9, 29, 58]$ ), a region previously associated with surprise coding in monkeys (Hayden et al., 2011), as well as in both insula and striatum (caudate nucleus) although these did not survive a stringent multiple comparisons correction (right insula:  $P = 0.16$ , peak  $t_{26} = 8.37$ ,  $[30, 26, -8]$ ; left insula:  $P = 0.19$ , peak  $t_{26} = 6.25$ ,  $[-33, 26, -5]$ ; left striatum ( $P = 0.84$ , peak  $t_{26} = 3.44$ ,  $[3, -25, -8]$ ). Posterior medial frontal cortex (pmPFC) and striatum are strongly implicated in the expression of a prediction error type signal in reinforcement learning (Pessiglione et al., 2006; Voon et al., 2010), as well as in signalling a discrepancy between an individual's behaviour and the behaviour of a group (Tomlin et al., 2013). An alternative measure of prediction error, where surprise was quantified as the discrepancy between the predicted choices of the partner and the partner's actual choices, did not yield significant activity in any area of the brain.



**Figure 4.6 Surprise as a mechanism underlying mPFC plasticity.** **A** Brain areas correlating with the surprise a subject experienced on observing a partner’s choice. **B** Correlation between the average surprise encoding in the striatum, extracted from ROI in (A) and  $[SN-SF]_{1-3}$  plasticity in mPFC. **C** Correlation between striatal surprise correlate and the shift in subjects’ discount rates towards the novel other. **A** is thresholded at  $P < 0.01$  uncorrected for visualization. a.u., arbitrary units.

A striatal prediction error type signal is known to drive learning through an influence on cortico-striatal plasticity (Reynolds and Wickens, 2002). In line with this notion, the BOLD correlate of the surprise about the partner’s choice in the striatum predicted the behavioural shift in subjects’ own discount rate (Figure 4.6C,  $r = 0.50$ ,  $P = 0.01$ ) as well as the change in self-to-novel versus change in self-to-familiar neuronal suppression over blocks in mPFC (Figure 4.6B,  $r = 0.41$ ,  $P = 0.04$ ). No such relationship was evident for pMFC or insula activity and mPFC plasticity ( $r = 0.04$ ,  $P = 0.84$  and  $r = 0.14$ ,  $P = 0.48$ , respectively).



**Figure 4.7 Mediation path diagram for discount rate shift as predicted from a striatal surprise signal.** The striatal surprise signal predicted  $[SN-SF]_{1-3}$  plasticity in the medial prefrontal cortex (path a) and the mediator (mPFC plasticity) predicted the shift of subjects’ own discount rate towards the discount rate of the novel other (path b, controlled for the striatal surprise signal). Importantly, there was a significant mediation effect (path  $a*b$ ), indicating that mPFC plasticity formally mediates the relationship between striatal surprise and the shift in discount rate. The direct path between striatal surprise and shift in discount rate, controlled for both mediators, was not significant (path c). The lines are labelled with path coefficients (SEs).

Finally, if prediction errors cause plasticity, and plasticity in turn causes the shift in subjects' discount rate, then plasticity in medial prefrontal cortex should formally mediate the impact of the striatal surprise signal on the shift in discount rate. We used single level mediation to test this hypothesis (Wager et al., 2008). The path model jointly tests three effects required if indeed mPFC plasticity provides the link between a surprise signal and the shift in discount rate: namely, the relationship between striatal surprise effects and mPFC plasticity (path a), the relationship between mPFC plasticity and shift in discount rate (path b), and a formal mediation effect (path a\*b) which indicates that each explains a part of the discount rate shift covariance while controlling for effects attributable to the other mediator. All three effects were significant in a mediation analysis (path a = 0.15, SE = 0.07, P = 0.04; path b = 0.30, SE = 0.12, P < 0.001; path a\*b = 0.05, SE = 0.03, P = 0.01, Figure 4.7) supporting the idea that prediction errors influence the discount rate by inducing medial prefrontal cortex plasticity, which in turn impacts upon choice behaviour. Hence, subjects with the largest striatal surprise signal at outcome of choice trials exhibited both the largest changes in representational similarity on probe trials, and the largest changes in preferences, suggesting a role for striatal prediction error signals in inducing cortical plasticity and associated behavioural change. A mediation analysis testing the opposite direction, i.e. a mediating effect of the striatal surprise signal on the relationship between mPFC plasticity and the shift in discount rate, was not significant (Supplementary Figure 4.5).

## 4.5 Discussion

The brain's representational architecture involves population codes wherein individual neurons contribute to a multitude of computations. We set out to investigate whether multiple neuronal representations can be updated simultaneously by learning-induced plasticity targeting one computation alone. We developed a novel approach that exploited repetition suppression (Grill-Spector et al., 1999; Henson et al., 2000) to probe the similarity between distinct neural representations (Barron et al., 2013), by interleaving probe valuation trials with decision blocks that induced prediction errors and learning. Whilst the biophysical mechanisms underlying fMRI repetition suppression remain ambiguous (Sobotka and Ringo, 1994), in a careful experimental design this approach allows inferences about population coding with respect to precise features of stimuli (Kourtzi and Kanwisher, 2001) or computations (Barron et al., 2013; Doeller et al., 2010).

I was interested in changes of value representational similarity over time. By asking subjects to evaluate presented options on behalf of themselves, a novel other whose preferences were acquired during on-line scanning and a familiar other whose preferences had previously been learnt, we could interrogate representational similarity in neuronal populations encoding valuation for these three agents. Learning about the preferences of a novel agent had a clear behavioural consequence evident in a shift in subjects' own, as well as their estimation of a familiar other's, discount rate. This behavioural effect coincided with an increase in neural representational similarity in the medial prefrontal cortex. This supports a view that value representations in the mPFC are not aligned to the frame of reference of an individual. Instead, an increase in neuronal overlap tied to a shift in behavioural preferences suggests that the mPFC encodes agent-independent representations of subjective value.

The presence of a learning-induced representational plasticity for value is likely to depend on generic learning mechanisms. The most influential computational account posits a role for a reward prediction error implemented via phasic activity of dopamine neurons (Schultz et al., 1997; Steinberg et al., 2013), a putative teaching signal for cortico-striatal learning (Calabresi et al., 2007; O'Doherty et al., 2004; Reynolds and Wickens, 2002). Prediction errors align with the dimension relevant for learning in a given situation. They manifest as a sensory prediction error when subjects learn to predict a sensory event (den Ouden et al., 2010), a probability prediction error when subjects learn about reward probability (Behrens et al., 2008) and a social expectancy prediction error when group preferences diverge from subjects' own valuations (Campbell-Meiklejohn et al., 2010; Klucharev et al., 2009). In the current experiment a prediction error, expressed in posterior medial frontal cortex (pmPFC), insula and striatum, corresponds to the surprise subjects experience when a partner's choice is incongruent with their own preference. This accords with previous studies demonstrating expression of a similar signal representing a discrepancy between one's own and a group's opinion (Berns et al., 2010; Campbell-Meiklejohn et al., 2010; Falk et al., 2010; Klucharev et al., 2009; Tomlin et al., 2013). Crucially, my results extend on these reports by showing this error coding is directly related to an expression of plasticity in mPFC, a region widely implicated in tracking preferences for stimuli (Lebreton et al., 2009), as well as inter-temporal preferences (Kable and Glimcher, 2007; Pine et al., 2009).

The mPFC region displaying the change in repetition suppression is a complex and heterogeneous area with strong connections to regions such as the amygdala, hippocampus,

hypothalamus and insula enabling access to sensory, visceral and emotional information. It is considered ideally placed for self-referential processing (Kelley et al., 2002; Magno and Allan, 2007) and for attributing value to stimuli across many reward contexts (Bartra et al., 2013; Clithero and Rangel, 2014), and internally generated states (Bouret and Richmond, 2010). However, a mPFC value computation is also remarkably flexible, and can occur even if direct experience is not available (Barron et al., 2013) or if there is a requirement for an abstract model of task structure (Hampton et al., 2006). This flexibility is vital in social cognition, where a model of the preferences and intentions of another individual needs to be decoupled from the physical and perceptual reality of a subject's own internal state (Mitchell, 2009; Nicolle et al., 2012). Traditionally, it has been suggested that such computations occur in distinct circuitries, where a ventral sector of the mPFC encoding subjective stimulus values (Boorman et al., 2009; O'Doherty, 2004; Plassmann et al., 2007) is complemented by a dorsal sector representing the mental states of others (Behrens et al., 2008, 2009; Frith and Frith, 2010; Yoshida et al., 2010). However, this notion is challenged by an observation that a dorsal-ventral axis can be better understood in terms of executed versus modelled choices (Nicolle et al., 2012). The latter observation supports the idea that the very same area encodes subjective value irrespective of the frame of reference, a notion strongly supported by our current observation that a behavioural shift towards the value of a novel agent is mirrored by an increase in neural overlap.

Irrespective of the exact nature of the observed plasticity, the underlying mechanism would seem to necessitate an overlap in neural populations encoding both a novel other, self and familiar other. How exactly might the brain calculate discounting preferences with neural populations that are prone to the observed shifts in preference? Theoretical studies suggest an agent's overall preferences might arise out of a summation over a distributed set of discounting units (Kurth-Nelson and Redish, 2009). This is consistent with recordings in rat orbitofrontal cortex demonstrating a distributed encoding of inter-temporal choice parameters across a neuronal population (Roesch et al., 2006). Similar gradients of discount factors have also been found in the human striatum (Tanaka et al., 2004) and medial prefrontal cortex (Wang et al., 2014). This suggests that some neuronal assemblies may represent a preference for fast discounting, favouring smaller-sooner returns, while others favour slow discounting. The discounting preference of each agent would be represented by population codes, implementing sets of weights over these discounting assemblies. The prediction errors a subject perceives when the novel other's choices differ from what they

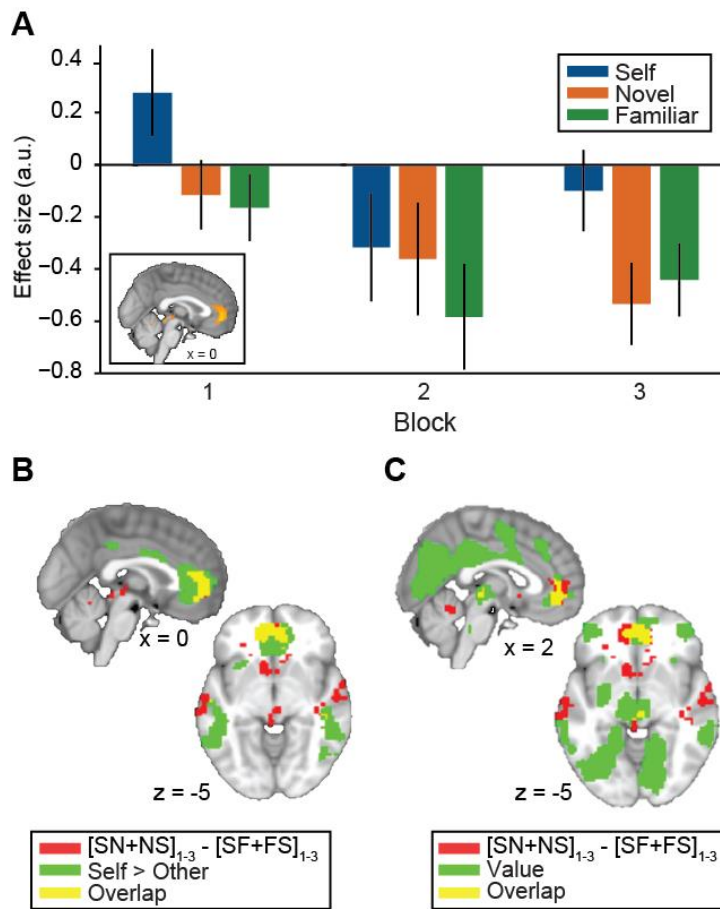
would have chosen for themselves could in principle change the weights within this pool, resulting in altered populations codes.

The fact that a common brain region is recruited when computing preferences for self and other might suggest that people initially draw on self-representations to make inferences about another person and only construct a novel representation through learning. Such a mechanism has been observed when constructing a representation for a novel good from a simultaneous activation of familiar components (Barron et al., 2013). However, this theory makes opposite neural predictions as it predicts repetition suppression at the beginning of the experiment as subjects draw on the same representation to choose for self and other. In this scenario a separate representation for a novel other is built over time and would predict disappearance of repetition suppression. Instead, we observe an increase in repetition suppression across time, an effect reminiscent of an increase in similarity between representations observed when subjects repeatedly evoke independent memories (Barron et al., 2013). Importantly, we can demonstrate this plasticity is not solely a neuronal phenomenon but also has profound behavioural consequences.

Note that subjects grow increasingly familiar with the novel other's preferences as the task progresses, whereas familiarity remains constant for the familiar other in the sense that there is no new learning in relation to this other. Since psychological constructs such as familiarity, but also similarity and physical proximity, have previously been demonstrated to upregulate mPFC activity (Jenkins et al., 2008; Krienen et al., 2010; Mitchell et al., 2006; Tamir and Mitchell, 2011), this raises the question whether an increase in familiarity might drive the plasticity effect. Importantly, our data is not consistent with such an account. First, activity for familiar and novel other does not differ in mPFC, not even at the beginning of the experiment, suggesting that the mPFC in our task does not respond to familiarity per se. Secondly, a mediation analysis suggests that it is a striatal surprise signal, the very opposite of familiarity, that drives the plasticity effect, which in turn drives the behavioural shift.

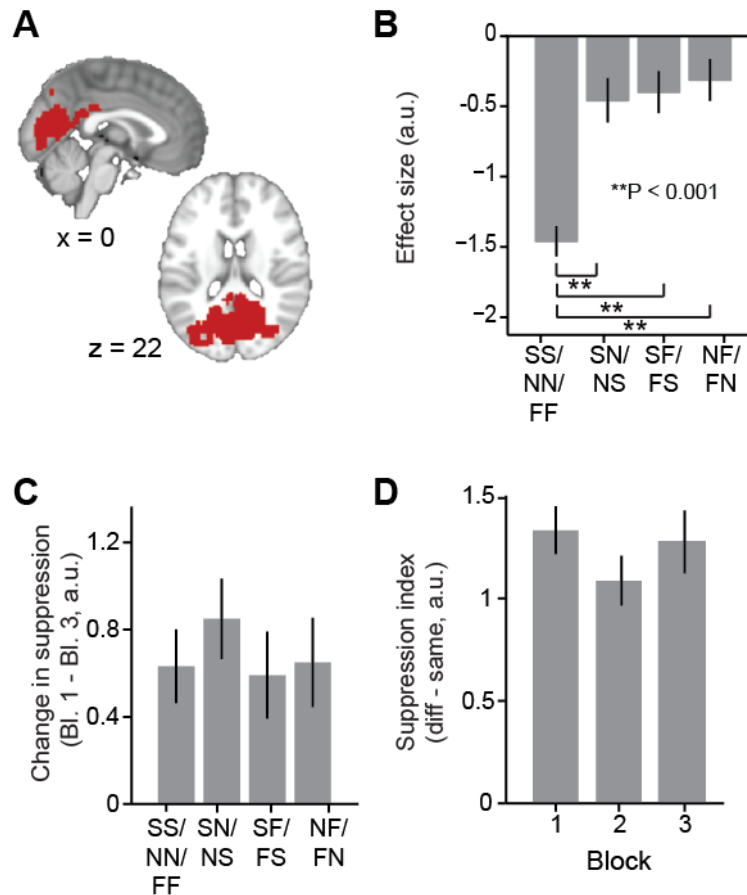
In conclusion, my data details a neuronal mechanism by which personal traits are directly susceptible to social influence. Such plasticity might be one of the key features underlying learning, because it allows for an integration of past experience when one has to extract information from novel samples. More broadly these findings pave the way for further studies of human social interactions at a more mechanistic level.

## 4.6 Supplementary Figures

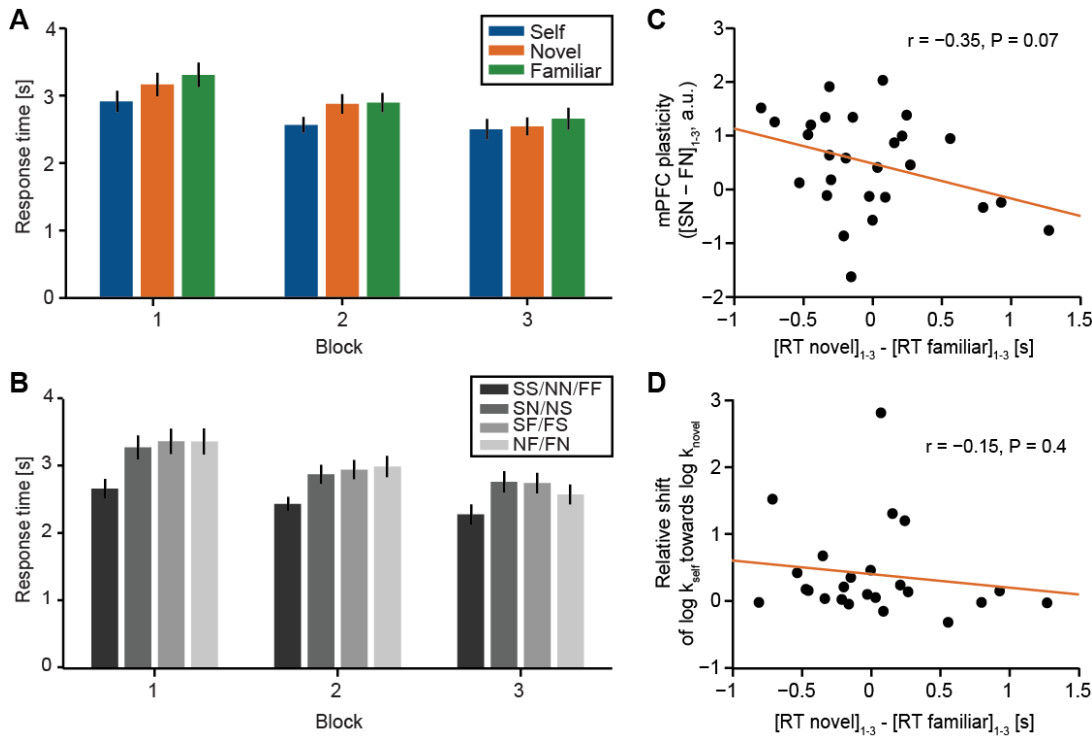


**Supplementary Figure 4.1 MPFC activity for self, other and value.** A Activity for self, novel and familiar other over blocks in the mPFC. A repeated measures ANOVA with within-subject factors “block” and “agent” showed that activity differed over blocks ( $F_{2,52} = 13.21$ ,  $P < 0.001$ ) and between agents ( $F_{2,52} = 4.19$ ,  $P = 0.05$ ). Furthermore, we found a block x agent interaction ( $F_{4,104} = 3.09$ ,  $P = 0.02$ ). Post-hoc tests revealed that activity in block 1 was different from activity in blocks 2 and 3 ( $P = 0.005$  and  $P < 0.001$ , respectively), but activity in blocks 2 and 3 did not differ ( $P = 0.88$ ). Activity between familiar and novel other did not differ ( $P = 1.0$ ), suggesting that the plasticity effect we report cannot be explained by differences in novelty/familiarity for the two agents. Parameter estimates were extracted from ROI shown in Figure 4.4 (see inset). B Overlap between self > other contrast and mPFC plasticity. Mean activity on self trials was higher than on other trials in left lateral parietal cortex ( $P < 0.001$ , peak  $t_{26} = 6.26$ , peak  $[-39, -79, 34]$ ) and in the mPFC ( $P < 0.001$ , peak  $t_{26} = 6.08$ , peak  $[9, 41, -5]$ ). Activity in the mPFC overlapped with the region showing an increase in suppression between self and novel, controlled for by an increase in suppression between self and familiar as depicted in Figure 4.4. The opposite contrast (other > self) only revealed activity in the visual cortex ( $P < 0.001$ , peak  $t_{26} = 8.83$ , peak  $[0, -94, 7]$ , not depicted). C Subjective value coding on probe trials and mPFC plasticity. Subjective value was encoded in left primary motor cortex ( $P < 0.001$ , peak  $t_{26} = 9.54$ , peak  $[-36, -25, 55]$ ), in right parietal cortex ( $P < 0.001$ , peak  $t_{26} = 5.04$ , peak  $[54 -16 22]$ ), in Brodmann area 10 ( $P = 0.031$ , peak  $t_{26} = 4.37$ , peak  $[-18, 62, 7]$ ) and in mPFC ( $P = 0.055$ , peak  $t_{26} = 4.26$ , peak  $[9, 44, 10]$ ). Activity in the mPFC overlapped with the region showing an increase in suppression between self and novel, controlled for by an increase in suppression between self and familiar as depicted in Figure 4.4. Results are reported at a cluster-defining threshold of  $P < 0.01$  uncorrected combined with a family-wise-error (FWE) corrected significance level of  $P < 0.05$ . All post-hoc tests were Bonferroni-corrected for multiple comparisons. a.u., arbitrary units.

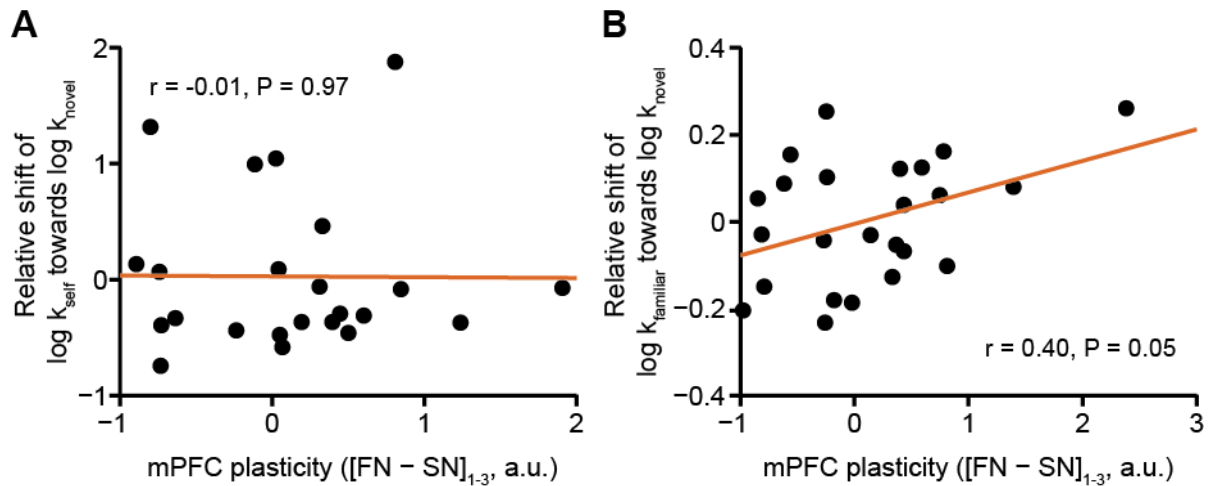




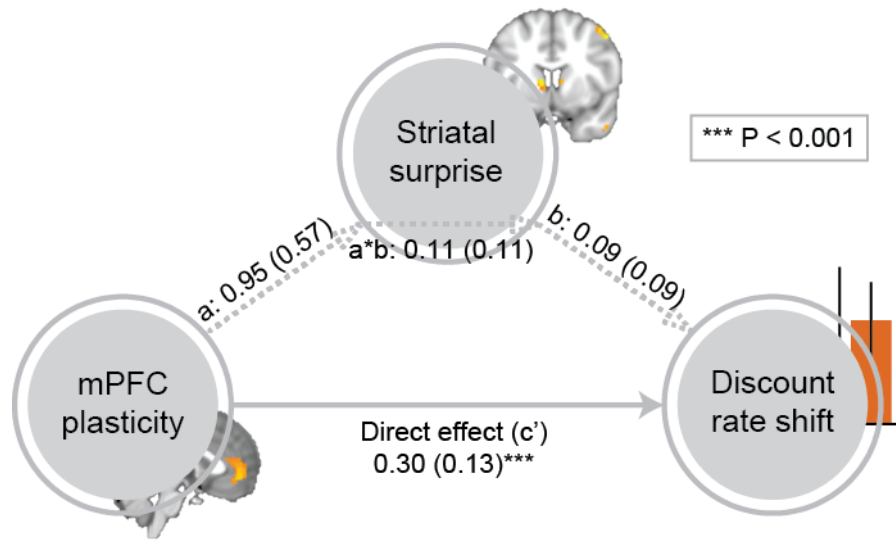
**Supplementary Figure 4.2 Repetition suppression in visual areas** **A** ROI used to interrogate plasticity effects in visual regions (thresholded at  $P < 0.001$  uncorrected for visualization). It was defined from a contrast indexing a main effect to any visual event in all three blocks. **B** Visual areas displayed significantly less activity when the agent from a preceding trial was repeated than when a different agent preceded a trial ( $F_{3,104} = 14.25$ ,  $P < 0.0001$ , block 1 only). **C** Suppression increased over blocks with no difference between conditions ( $F_{3,104} = 0.37$ ,  $P = 0.78$ ). **D** The difference in mean activity on same-agent-preceding trials versus different-agent-preceding trials did not change over blocks ( $F_{2,78} = 0.32$ ,  $P = 0.73$ ). SS: self-preceded-by-self, NN: novel-preceded-by-novel, FF: familiar-preceded-by-familiar, SN: novel-preceded-by-self, NS: self-preceded-by-novel, SF: familiar-preceded-by-self, FS: self-preceded-by-familiar, NF: familiar-preceded-by-novel, FN: novel-preceded-by-familiar. a.u., arbitrary units.



**Supplementary Figure 4.3 Response time analyses.** **A** Response times for self, novel and familiar other probe trials over blocks. A repeated measures ANOVA with within-subject factors “block” and “agent” showed different response times between blocks ( $F_{2,52} = 21.28$ ,  $P < 0.001$ ), agents ( $F_{2,52} = 19.8$ ,  $P < 0.001$ ) as well as a block x agent interaction ( $F_{4,104} = 2.88$ ,  $P = 0.03$ ). Post-hoc tests reveal differences between all blocks (blocks 1/2:  $P = 0.002$ , blocks 1/3:  $P < 0.001$ , blocks 2/3:  $P = 0.04$ ). Furthermore, subjects respond faster for self than for other ( $P < 0.001$  for self/novel and self/familiar). Importantly, we found no significant difference in response times for novel and familiar other ( $P = 0.29$ ), confirming that there is no novelty/familiarity effect. **B** To test for behavioural suppression effects that are in line with the neural suppression effects, we performed a repeated measures ANOVA with Greenhouse-Geisser correction with within subject factors “block” and “suppression condition” (SS/NN/FF, SN/NS, SF/FS, FN/NF) on subjects’ response times on probe trials. We found a main effect of block ( $F_{1,4,37.5} = 21.3$ ,  $P < 0.001$ ), a main effect of condition ( $F_{2,3,60.6} = 41.7$ ,  $P < 0.001$ ), and a block x condition interaction ( $F_{4,5,118.8} = 4.5$ ,  $P = 0.001$ ). Post-hoc paired tests revealed that SS/NN/FF differed from all other conditions ( $P < 0.001$ ), but SN/NS, SF/FS and FN/NF did not differ from each other (all comparisons  $P = 1.0$ ). This emphasizes that the neural suppression effects between self and novel, and between self and familiar, respectively, cannot simply be explained by faster processing speed. **C** Correlation between response time facilitation for the novel other ( $\text{RT novel block 1} - \text{RT novel block 3} - (\text{RT familiar block 1} - \text{RT familiar block 3})$ ) and  $[\text{SN} - \text{SF}]_{1,3}$  plasticity effect in mPFC (ROI from Figure 4.4A). Response time facilitation, a crude index of increasing familiarity, shows a trend towards a negative correlation with the neural plasticity effect ( $r = -0.35$ ,  $P = 0.07$ ). This is in line with an observation that the opposite of familiarity, namely surprise, is a better predictor of mPFC plasticity. **D** Correlation between response time facilitation for the novel other and behavioural discount rate shift of self towards the novel other ( $r = -0.15$ ,  $P = 0.4$ ). All post-hoc tests were Bonferroni-corrected for multiple comparisons. SS: self-preceded-by-self, NN: novel-preceded-by-novel, FF: familiar-preceded-by-familiar, SN: novel-preceded-by-self, NS: self-preceded-by-novel, SF: familiar-preceded-by-self, FS: self-preceded-by-familiar, NF: familiar-preceded-by-novel, FN: novel-preceded-by-familiar. a.u., arbitrary units.



**Supplementary Figure 4.4 Relationship between [FN-SN]1-3 plasticity and shift in discount rate.** **A** Partial correlation between the change in suppression between familiar and novel other controlled for by the change in suppression between self and novel other and the shift of subjects' own discount rate towards the novel other. **B** Partial correlation between the change in suppression between self and novel other controlled for by the change in suppression between self and familiar other and the shift of subjects' estimate of the familiar other's discount rate towards the novel other. Parameter estimates in A and B were extracted from the mPFC ROI shown in Figure 4.4. To account for the correlation between subjects' own shift in discount rate and the shift in their estimate of the familiar other's discount rate, we performed partial correlations, i.e. the familiar shift was removed from both signals in A and the self shift was removed from both signals in B. These analyses indicate that the change of familiar-to-novel suppression versus the change in self-to-novel suppression predicted a shift in subjects' estimate of the familiar other's discount rate (partial correlation,  $r = 0.40$ ,  $P = 0.05$ , B) but not the shift in subjects' own discount rate (partial correlation,  $r = -0.01$ ,  $P = 0.97$ , A). This emphasizes the relationship between increasing neuronal similarity between two agents' value representations and increasing behavioural similarity, as depicted in Figure 4.5. However, note that these data are merely suggestive, as removal of the rightmost data point in B affects the significance of the result. a.u., arbitrary units.



**Supplementary Figure 4.5 Mediation path diagram for discount rate shift as predicted from the mPFC plasticity signal.** The mPFC plasticity did not predict the striatal surprise about the other's choice (path a,  $p = 0.054$ ), and the mediator (striatal surprise) did not predict the shift of subjects' own discount rate toward the discount rate of the novel other (path b, controlled for the mPFC plasticity signal,  $p = 0.2$ ). There was also no significant mediation effect (path ab,  $p = 0.17$ ), demonstrating that striatal surprise does not formally mediate the relationship between mPFC plasticity and the shift in discount rate. However, the direct path between the mPFC plasticity signal and the shift in discount rate, controlled for both mediators, was highly significant (path c',  $p = 0.0007$ ). The lines are labeled with path coefficients (SEs).

## **5 A MAP OF ABSTRACT RELATIONAL KNOWLEDGE IN HUMAN ENTORHINAL CORTEX**

\* This chapter is prepared for publication as the following article:

Garvert MM, Dolan RJ and Behrens TEJ. A map of abstract relational knowledge in human entorhinal cortex.

## 5.1 Abstract

The hippocampal-entorhinal system encodes a map of the relationships between landmarks in space that is used in spatial navigation. Goal-directed behaviour outside of spatial navigation similarly requires a neural representation of the relationship between objects, events and other types of information, and such abstract forms of relational knowledge rely on the same neural system. It is not known whether such abstract information can profit from organisational principles that govern spatial relationships. Here, we use human fMRI adaptation to show that the entorhinal cortex can represent metric relationships between abstract objects. These metrics enable us to reconstruct a simple map-like knowledge structure directly from the entorhinal BOLD signal in a situation where relationships are non-spatial rather than spatial, discrete rather than continuous and unavailable to conscious awareness.

## 5.2 Introduction

Animals efficiently extract abstract relationships between landmarks, events and other types of conceptual information, often from limited experience. Knowing such regularities can help us act in an environment, because the relationships between items that have never been experienced together can easily be computed and exploited for making novel inference. In physical space, spatially tuned cells in the hippocampal-entorhinal system have precise place (O'Keefe and Dostrovsky, 1971) and grid (Hafting et al., 2005) codes that may form the neural basis of a 'cognitive map' (O'Keefe and Nadel, 1978). It is likely that the particular form of these representations enables rapid computations of critical features of spatial relationships such as distances and vector paths (Bush et al., 2015; Stemmler et al., 2015). The potential for such rapid online computations embedded into the neuronal representations may explain how animals can find novel paths through space (McNaughton et al., 2006; Mittelstaedt and Mittelstaedt, 1980) or rapidly reroute when obstacles are introduced (Alvernhe et al., 2011) or removed (Alvernhe et al., 2008). Indeed, in humans, signals that encode distance metrics between landmarks (Howard et al., 2014; Morgan et al., 2011), and directions to goals (Chadwick et al., 2015) can be read out directly from fMRI data in entorhinal cortex.

The hippocampal-entorhinal system can also encode a metric of non-spatial relationships if these are directly analogous to physical space. For example, the hippocampus encodes the relationship between stimuli varying along a one-dimensional continuous scale such as time (Ezzyat and Davachi, 2014) as well as the angles between locations in an abstract two-dimensional stimulus space, where both stimulus dimensions vary along a continuous scale (Tavares et al., 2015). In the entorhinal cortex, a hexagonally-symmetric code can be observed when humans navigate two-dimensional conceptual knowledge, suggesting that grid cells may provide a relational code for non-spatial relational information varying along two continuous, abstract dimensions (Constantinescu et al.). However, most relational problems that are of fundamental importance to memory and cognition cannot be mapped onto one or two continuous axes. Instead, relationships between stimuli are often high-dimensional, discrete rather than continuous, and unavailable to self-reported awareness. For example, transitive inference, an essential component of intelligent reasoning across species and cognitive domains, typically requires a model of the discrete relationships between stimuli in our environment (Tervo et al., 2016).

Notably, the hippocampus also supports the formation of associations between non-spatial and discrete stimuli across arbitrary stimulus dimensions, such as temporal co-occurrence (Schapiro et al., 2012, 2013) or social rank (Kumaran et al., 2012) and organizes behaviourally relevant stimulus categories in a hierarchy (McKenzie et al., 2014). The hippocampus also generalizes over individual episodes (Komorowski et al., 2013) and combines newly formed associations between discrete stimuli to enable transitive inference (Collin et al., 2015; Heckers et al., 2004; Horner et al., 2015; Preston et al., 2004; Schlichting et al., 2015), similar to the computation of novel paths in physical space. Associations can also directly influence behaviour in novel decision making situations by allowing value to spread across associated stimulus representations (Wimmer and Shohamy, 2012).

While the hippocampal-entorhinal system therefore clearly forms relational codes for discrete events which enables cognitive flexibility across non-spatial domains, it is unknown what metric underlies this organization and whether the computations are the result of an explicit map-like representation of abstract relationships where distances between discrete stimuli are preserved. If the representation of non-spatial relational information relies on the same neural algorithms as the representation of physical space (Buzsáki and Moser, 2013), then non-spatial and discrete relational information might share the neural codes organizing spatial relationships in a map and non-spatial relationships could profit from the same

organizational principles that govern physical space. Notably, a map-like representation of abstract relational knowledge would also provide new constraints on the computations that the hippocampal-entorhinal system is likely to perform in non-spatial reasoning.

Here, we explicitly tested this notion using a functional magnetic resonance imaging (fMRI) adaptation paradigm that allowed us to quantify the relationships between neuronal object representations in a neuronal representational space following an implicit learning paradigm. We presented human participants with sequences of objects where stimulus transitions were drawn from random walks along a graph structure. Within entorhinal cortex, a map-like organisation of the relationships between object representations could be extracted from functional magnetic resonance imaging (fMRI) adaptation data acquired on the subsequent day. This suggests that the brain automatically organizes abstract relational information in a map even if the relationships between objects are non-spatial rather than spatial, discrete rather than continuous and unavailable to conscious awareness.

## 5.3 Methods

### 5.3.1 Subjects

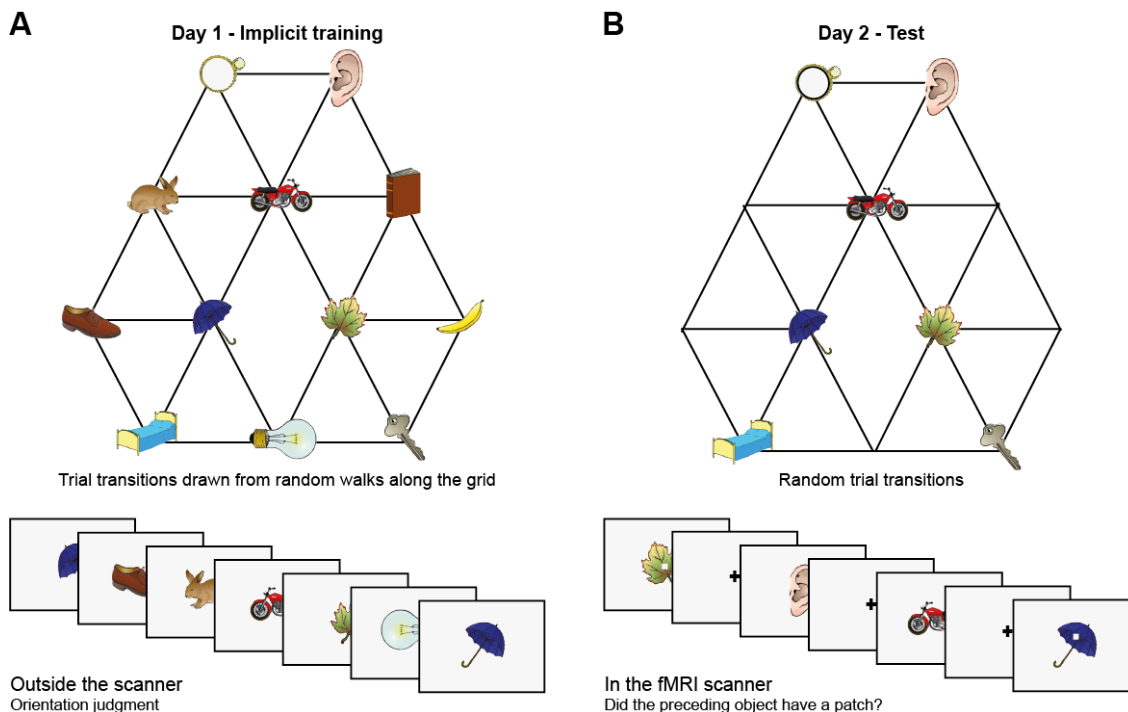
23 volunteers (aged 18-31, mean age  $\pm$  std  $23.5 \pm 3.7$  years, 15 males) with normal or corrected-to-normal vision and no history of neurological or psychiatric disorder participated in this experiment. All subjects gave written informed consent and the study was approved by the University College London Hospitals Ethics Committee. The study took place at the Wellcome Trust Centre for Neuroimaging. Subjects were naïve to the purpose of the experiment.

### 5.3.2 Stimuli and task

31 coloured and shaded object images which were similar in terms of their familiarity and complexity were selected from the “Snodgrass and Vanderwart ‘Like’ Objects’ picture set (<http://wiki.cnhc.cmu.edu/Objects>, Rossion and Pourtois, 2004). For each subject, a subset of 12 objects was chosen and randomly assigned to the 12 nodes of the graph shown in Figure 5.1A. On day 1, subjects were exposed to objects sequences generated from a random walk on the graph, where only objects that were directly connected to an object by a link could follow a presentation of this object. To avoid local repetitions we constrained



sequences such that at least 3 objects had to occur between any two presentations of the same object. Each object was randomly presented in one of two orientations, which were mirror images of each other.



**Figure 5.1 Experimental design.** **A** Graph structure, used to generate stimulus sequences on day 1. Trial transitions were drawn from random walks along the graph. **B** Objects on reduced graph, presented to subjects in the scanner on day 2. Trial transitions were random. In both sessions, participants performed simple behavioural cover tasks.

Before the start of the experiment, subjects were shown the entire set of 12 stimuli and instructed to remember which of two buttons to press for a particular object orientation. During the actual training, subjects were instructed to press the button associated with the stimulus orientation while watching the objects sequences as quickly and accurately as possible. Visual feedback after each button press indicated whether a response was correct. Key assignment was counterbalanced across subjects. Subjects learnt to perform the task quickly and accurately (Supplementary Figure 5.1). Stimuli were presented for 2 s and each experimental block consisted of 133 object presentations. Subjects performed this experiment for 12 blocks in total. Between blocks (ca. every 5 min), they were free to take self-paced breaks.

On the next day, subjects were presented with object sequences in the scanner. Only the 7 objects corresponding to the locations illustrated in Figure 5.1B were presented to reduce the total number of stimulus-stimulus transitions and thereby increase statistical power for

our key question of interest. Furthermore, stimulus transitions did not follow the graph structure, but were instead randomized with a constraint that each of the 42 possible object transitions occurred exactly 10 times per block (objects were never repeated).

The fMRI experiment consisted of 421 items per run and 3 experimental runs. Stimuli were presented for 1 s, with a jittered inter-trial interval (ITI) generated from a truncated Poisson distribution with a mean of 2s. While observing the object sequences subjects performed a cover task of infrequently reporting by button press whether a little grey patch had occurred on a preceding trial. The patch was present on a randomly selected 50% of the objects. Trials on which subjects had to report the existence of a grey patch were signalled by a green cross during the inter-stimulus-interval instead of the standard white cross. The cross was green exactly once after each of the 42 possible transitions, i.e. in 10% of the total number of trials. In 50% of those cases, a patch had been present on the preceding trial. Each correct button press was rewarded with £0.10 paid out in addition to a £33 show-up fee to ensure that subjects attended to the stimuli. Subjects received a brief training on this task before they performed it in the scanner. Key assignment was counterbalanced across subjects. Subjects performed the cover task very well (correct performance rate across subjects:  $94\% \pm 3\%$ , mean  $\pm$  s.e.m.) confirming that they paid attention to the presented objects throughout the duration of the scan.

### 5.3.3 fMRI data acquisition and pre-processing

Visual stimuli were projected onto a screen via a computer monitor. Subjects indicated their choice using an MRI-compatible button box.

MRI data was acquired using a 32-channel head coil on a 3 Tesla Allegra scanner (Siemens, Erlangen, Germany). A T2\*-weighted echo-planar (EPI) sequence was used to collect 43 transverse slices (ascending order) of 2 mm thickness with 1-mm gap and in-plane resolution of 3x3 mm, a repetition time of 3.01 s and an echo time of 70 ms. Slices were tilted by 30° relative to the rostro-caudal axis and a local z-shim with a moment of -0.4mT/m was applied to the OFC region (Weiskopf et al. 2006). The first five volumes of each block were discarded to allow for scanner equilibration. After the experimental sessions, a T1-weighted anatomical scan with 1x1x1 mm resolution was acquired. In addition, a whole-brain field map with dual echo-time images (TE1 = 10 ms, TE2 = 14.76 ms, resolution 3x3x3 mm) was obtained to measure and later correct for geometric distortions due to susceptibility-induced field inhomogeneities.

We performed slice time correction, corrected for signal-bias and realigned functional scans to the first volume in the sequence using a six-parameter rigid body transformation to correct for motion. Images were then spatially normalized by warping subject-specific images to an MNI reference brain, and smoothed using an 8 mm full-width at half maximum Gaussian kernel. All pre-processing steps were performed with SPM12 (Wellcome Trust Centre for Neuroimaging, <http://www.fil.ion.ucl.ac.uk/spm>).

#### 5.3.4 fMRI data analysis

We implemented two event-related general linear models (GLM) to analyse the fMRI data. The first GLM contained separate onset regressors for each of the 7 objects with a patch, and without a patch. Each onset regressor was accompanied by a parametric regressor indicating the distance between the object on trial  $t$  and the preceding object on trial  $t-1$  on the graph presented in Figure 5.1B (distance 1, 2 or 3). Furthermore, a button press regressor was included as a regressor of no interest. Trials associated with a button press, and the two subsequent trials were not included in the main regressors to avoid button-press related artefacts. All regressors were convolved with a canonical hemodynamic response function. Because of the sensitivity of the BOLD signal to motion and physiological noise, we included six motion regressors obtained during realignment as well as ten regressors for cardiac phase, six for respiratory phase and one for respiratory volume extracted with an in-house developed Matlab toolbox as nuisance regressors (Hutton et al., 2011). Models for cardiac and respiratory phase and their aliased harmonics were based on RETROICOR (Glover et al., 2000). Sessions were modelled separately within the GLM.

In the second GLM, all 42 possible object transitions (object 1 preceded by object 2; object 1 preceded by object 3, ..., object 7 preceded by object 6) were modelled separately for patch trials and no-patch trials. Furthermore, button presses were included as a regressor of no interest. All regressors were convolved with a canonical hemodynamic response function. Again, motion and physiological noise were regressed out by including the six motion regressors obtained during realignment as well as ten regressors for cardiac phase, six for respiratory phase and one for respiratory volume as nuisance regressors (Hutton et al., 2011).

To investigate whether activity scales with distance on the graph in a whole-brain analysis, we assessed the effect of the parametric distance modulator for the non-patch trials in GLM 1. The contrast images of all subjects from the first level were analysed as a second-level

random effects analysis. We report our results in the entorhinal cortex, as this was our *a priori* region of interest, at a cluster-defining statistical threshold of  $p < 0.001$  uncorrected, combined with small volume correction (SVC) for multiple comparisons (peak-level family-wise error (FWE) corrected at  $p < 0.05$ ). For the SVC procedure we used two different anatomical masks. The first mask consisted of the entorhinal cortex and subiculum alone and was received with thanks from Martin Chadwick (Chadwick et al., 2015, Supplementary Figure 5.2A). The second mask also contained other medial temporal lobe regions implicated in encoding physical space and comprised hippocampus, entorhinal cortex and parahippocampal cortex as defined according to the maximum probability tissue labels provided by Neuromorphometrics, Inc (Supplementary Figure 5.2B). Activations in other brain regions were only considered significant at a level of  $P < 0.001$  uncorrected if they survived whole brain FWE correction at the cluster level ( $P < 0.05$ ). While no areas survived this stringent correction for multiple comparisons, other regions are reported in the Supplementary Material at a threshold of  $P < 0.01$  uncorrected for multiple comparisons for completeness (Supplementary Figure 5.3).

To independently test a distance-dependent scaling of activity within entorhinal cortex we defined two different regions of interest (ROIs) based on GLM 2. The first region of interest was defined on the basis of decreased activity in transitions where the preceding object was directly connected with the current object (e.g. regressors corresponding to transition 1-2, 6-4 or 5-7, see Supplementary Figure 5.1C) relative to all other transitions (e.g. regressors corresponding to transition 4-2, 7,4 or 1-7, i.e. [non-connected - connected]). This contrast revealed that clusters in bilateral entorhinal cortex show more adaptation if a preceding object is connected with a currently presented object, relative to a situation where the preceding objects is two or three links away (green, Figure 5.2B, and Supplementary Figure 5.3C). This defined a region of interest (thresholded at a p-value of 0.01) from which we then extracted parameter estimates for each of the 42 no-patch transitions and tested for an orthogonal distance effect, namely whether activity differed for transitions with distance 2 relative to transitions of distance 3 using a two-sided t-test.

In a second, independent, test we defined a bilateral entorhinal ROI based on the following contrast: [transitions with 3 links between the relevant objects] - [transitions with 2 links between the relevant objects] for non-patch trials. This contrast is orthogonal to the first contrast, and identified brain regions that responded more strongly on a trial if the preceding object was 3 links, rather than 2 links away (red, Figure 5.2B and Supplementary

Figure 5.3). Again, we extracted parameter estimates for each of the 42 non-patch onset regressors from this ROI, and performed an orthogonal test for distant-dependent scaling by investigating whether activity in this region was also significantly different for connected vs. non-connected objects (e.g. transition 1-2, 6-4 or 5-7 versus transition 4-2, 7,4 or 1-7, Supplementary Figure 5.1C).

Note the distance-dependent scaling effects cannot be explained by object-specific differences in activity within these ROIs. While the mean activity for different objects differs slightly, but non-significantly (Supplementary Figure 5.4A,C,E), removing these main effects by subtracting the mean activity for each object before performing the above described analyses does not alter the results (Supplementary Figure 5.4B,D,F).

In a further independent test of a distance-dependent scaling of activity in the entorhinal cortex, we extracted parameter estimates from a region of interest defined based on an independent study investigating the representation of a geocentric goal direction in the entorhinal/subicular region (Chadwick et al., 2015). Specifically, we extracted parameter estimates for the 42 non-patch transitions from the peak voxel reported in his study (MNI coordinates:  $[-20, -25, -24]$ ). This definition of a region of interest was non-biased, and allowed us to test directly test for a distance-dependent scaling of activity. We first performed a one-way ANOVA on the parameter estimates sorted according to distance (Figure 5.3E). To investigate whether information is organized with respect to the distance relationship or with respect to the average time that passed between the occurrence of two objects during training, we performed a multiple linear regression. In this regression analysis, we included one regressor denoting the distance between object pairs (1, 2 and 3) and second regressor accounting for the number of objects that had occurred between any pair of objects  $i$  and  $j$  during training. Since the duration of object presentations and the ITI during training were constant, this measure was directly proportional to the time elapsed between the occurrence of the two objects. The dependent variable in the regression analysis was the neural activity for the 42 non-patch transition regressors extracted from this independently defined peak voxel (Figure 5.3A). To test for the directionality of the distance effect in the entorhinal cortex, we exploited the fact that subjects were not exposed to object transitions in the two directions (e.g. 5 followed by 3 vs. 3 followed by 5) equally often. In fact, across subjects there was a large variability in the absolute difference between the number of times one direction vs. the other direction were experienced during training (Figure 5.3B). To test for non-directionality in the neural signature, a feature of a map-like structure, we converted the

number of times each transition was experienced into a distance measure for each individual subject according to the following equation:

$$d = 1 - \frac{c}{1 + c_{\max}} \quad (5.1)$$

Here,  $d$  denotes the length of the shortest path between two connected objects. It is computed based on the number of times this particular transition was experienced during training ( $c$ ) relative to the number of times the transition that was visited most often during training was experienced ( $c_{\max}$ ). The length of the path between objects that were two or three links away was then computed as the single-source shortest path between these objects (by adding the pathlengths for connected objects linking these two objects and choosing the shortest one). To compute the ‘symmetric shortest path measure’, the directional path-length measures (e.g. 5-2 and 2-5) were averaged. The directional, and the symmetric shortest path measures were used as regressors to predict the neural signal extracted from the peak voxel (Figure 5.3C).

To visualize the representation of the graph structure in the entorhinal cortex, we performed multi-dimensional scaling (MDS) on the neural activity extracted from the same peak voxel. MDS arranges objects spatially such that the distances between them in space correspond to their similarities as defined by the distance matrix as well as possible. Here, we estimated the configuration of objects in two dimensions using the corresponding inbuilt matlab function. Specifically, MDS was performed on a matrix denoting the mean neural activity across subjects for each pair of transitions. For example, element 2-5 in the matrix corresponded to the average activity across subjects on trials where object 5 was preceded by object 2 and element 5-2 corresponded to the average activity across subjects on trials where object 2 was preceded by object 5. Because neural activity scales with distance, this matrix effectively corresponds to a distance or similarity matrix. Note that multi-dimensional scaling can only be performed on symmetric matrices with positive entries. We therefore normalized the matrix by subtracting the minimum value of the matrix and adding 1, and then symmetrized it by averaging the top and the bottom triangles.

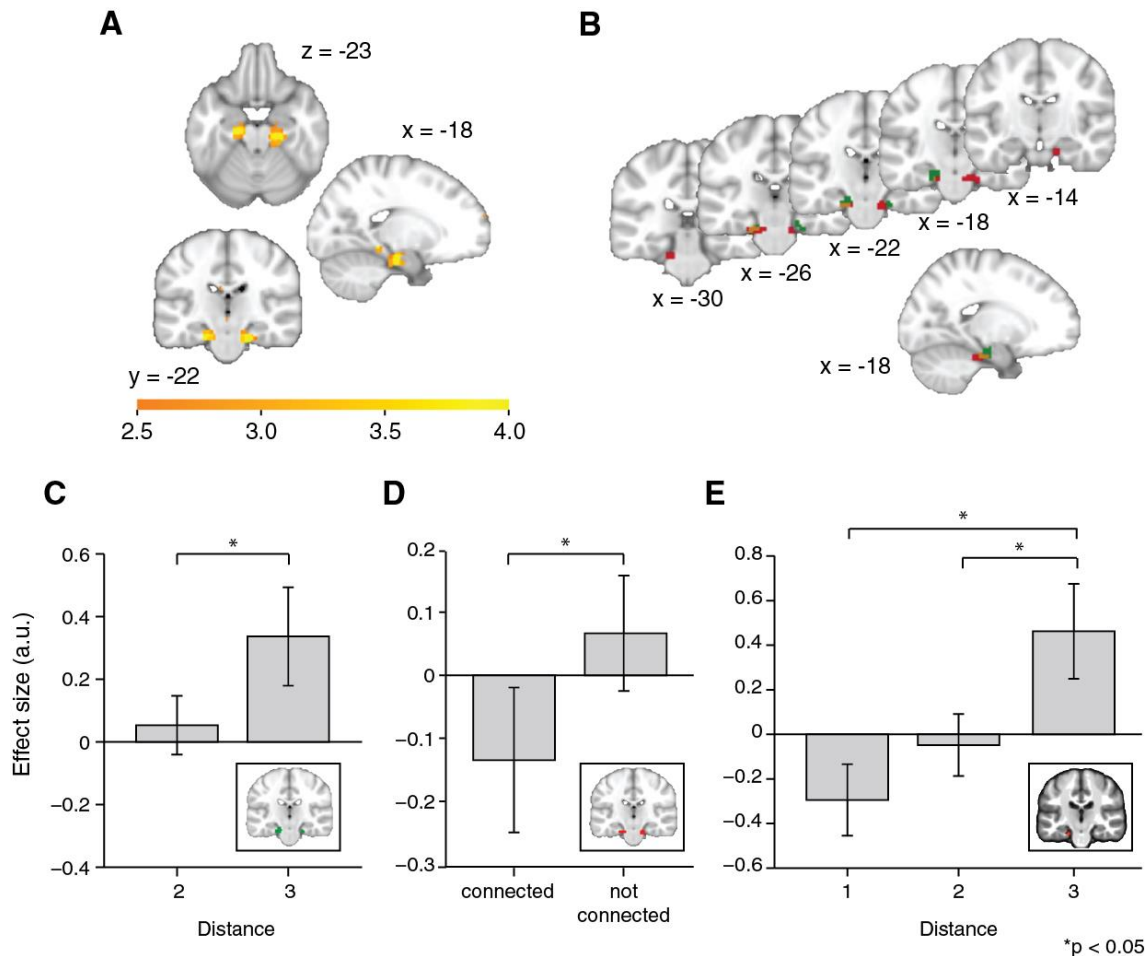
To test for a specificity in the distance effect, we also extracted parameter estimates from an ROI located in the visual cortex defined based on a main effect of object onset across blocks (thresholded at  $t > 6$ ) and performed the same analyses as in the entorhinal cortex (Supplementary Figure 5.3F).

## 5.4 Results

We first exposed 23 human participants to object sequences whose stimulus transitions, unbeknownst to them, were determined by a random walk in a graph (Figure 5.1A). Subjects performed a behavioural cover task, in which they learned to associate a random stimulus orientation with a button press. In the task instructions, any reference to a sequence or an underlying structure was avoided. After the fMRI experiment subjects were debriefed and none reported any explicit knowledge of structure in the task. To test whether this exposure to object sequences induced implicit knowledge about the graph, we scanned the subjects on a subsequent day using fMRI while exposing them to a subset of the same objects presented in a random order (only a reduced graph was presented to increase statistical power, Figure 5.1B). In 1/10th of the fMRI trials, subjects performed an unrelated cover task, reporting whether a grey patch was present on the screen. Neither accuracy nor reaction time in this task depended on the object on screen or the transition structure (Supplementary Figure 5.1).

We exploited fMRI adaptation (Grill-Spector et al., 2006) to investigate the representational similarity for different objects on the graph. We reasoned that in regions encoding a map-like representation of the overall task structure, the degree of similarity in neural representation, and therefore the fMRI adaptation (Kourtzi and Kanwisher, 2001), should decrease as a function of distance between items on the graph. Based on this reasoning we first looked for brain regions whose fMRI response to each object increased as a linear function of the graph-distance of the preceding item. We focused our analysis on the hippocampal-entorhinal system, as this medial temporal lobe region is considered the substrate for encoding maps of space.

This adaptation analysis revealed a peak bilaterally in the entorhinal cortex (Figure 5.2A, family-wise error-corrected at peak level within a bilateral entorhinal cortex/subiculum mask, left  $P = 0.014$ , peak  $t_{22} = 3.78$ ,  $[-18, -19, -22]$  and right  $P = 0.006$ , peak  $t_{22} = 4.75$ ,  $[24, -25, -22]$ ). A right, but not the left, peak also survived SVC for a larger region of interest (ROI) comprising the hippocampus, parahippocampal cortex and entorhinal cortex, left  $P = 0.058$  and right  $P = 0.026$ , see ROIs in Supplementary Figure 5.2).



**Figure 5.2 fMRI adaptation in the entorhinal cortex decreases with distance on the graph. A** Whole-brain analysis showing a decrease in fMRI adaptation with graph distance in the entorhinal cortex, thresholded at  $P < 0.01$ , uncorrected for visualization. **B** Within entorhinal cortex, green indicates greater suppression if the preceding stimulus was a neighbour relative to a stimulus two or three links away. Red indicates greater suppression if a preceding stimulus was two links away than three links away. The depicted areas were used as regions of interest for analyses in **C** (green) and **D** (red). **C** Parameter estimates for link 2 vs link 3 transitions extracted from the green entorhinal ROI in **B**. Other brain areas do not show this increase in activity with distance (Supplementary Figure 5.3). **D** Parameter estimates extracted from the red entorhinal ROI in **B**, sorted according to whether objects were connected on the graph or not. **E** Parameter estimates extracted from the peak MNI coordinate reported in Chadwick et al. (2015),  $[-20, -25, -24]$  and sorted according to distance. Error bars show mean and s.e.m.; a.u., arbitrary units.

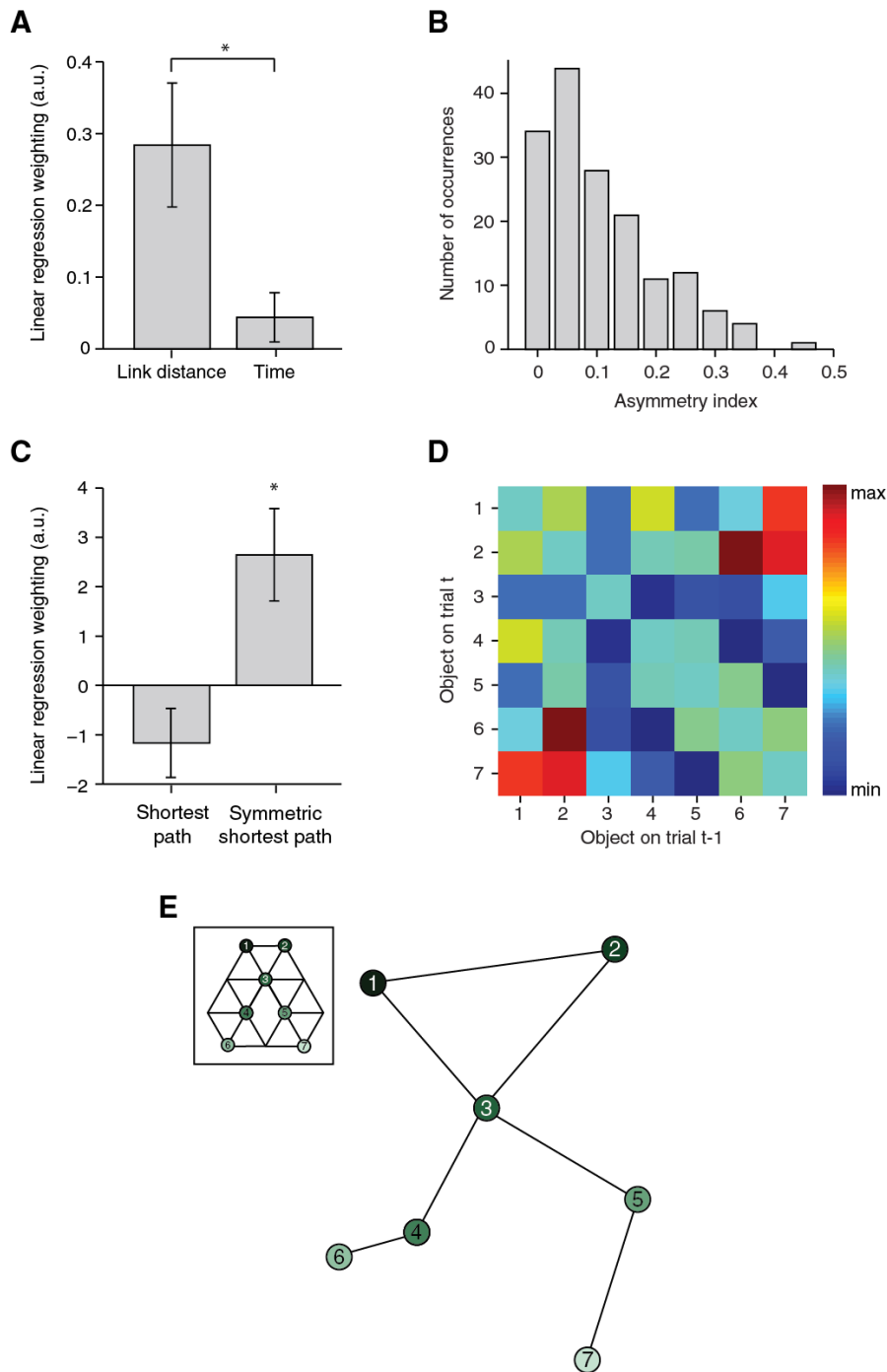
To confirm the statistical robustness of the effect, and to test whether it reflected a gradual increase with distance, we separated the effect into two orthogonal components. These components comprised the difference between connected links (length 1) and all other transitions (lengths 2,3; Figure 5.2B, green), and the difference between transitions of length 2 and those of length 3 (Figure 5.2B, red). These two independent contrasts were used to define ROIs bilaterally in overlapping regions of the entorhinal cortex (both thresholded at  $p < 0.01$  uncorrected; contrast 1: left peak  $t_{22} = 3.71$ ;  $[-21, -22, -25]$  and right peak  $t_{22} = 3.99$ ;



[21, -28, -22], contrast 2: left peak  $t_{22} = 3.21$ , [-12, -16, -25] and right peak  $t_{22} = 4.04$ , [21 -19 -28]). Because of their statistical independence, we could use the ROI from one contrast to extract data for the corollary test ( $t_{22} = 2.27$ ,  $p = 0.03$  for length 2 vs. length 3 in ROI 1, Figure 5.2C; and  $t_{22} = 2.21$ ,  $P = 0.04$  for connected vs. all other links in ROI 2, Figure 5.2D). This pair of tests suggest that the fMRI adaptation faithfully represents the link distance. These tests obviate questions of multiple comparisons, because in each case the data is selected from one contrast, and an orthogonal contrast was used for the test statistic.

To further demonstrate this within a single test, we used a peak location taken from an independent study investigating spatial maps (Chadwick et al., 2015). Extracting data from this coordinate (ROI 3) revealed a linear effect of graph distance (Figure 5.2E,  $F_{2,44} = 10.04$ ,  $p < 0.001$ ), and correspondingly a significant difference between distances of lengths 1 and 3 ( $t_{22} = 3.71$ ,  $P = 0.001$ ) and lengths 2 and 3 ( $t_{22} = 3.19$ ,  $P = 0.004$ ), but not between distances of lengths 1 and 2 ( $t_{22} = 1.67$ ,  $p = 0.11$ ).

Although this distance effect is suggestive of a map-like organisation, it might also merely reflect the temporal proximity between two objects during training. In a direct comparison of the temporal versus distance relationship between pairs of objects, the number of links ( $t_{22} = 3.29$ ,  $P = 0.003$ ), but not time ( $t_{22} = 1.27$ ,  $P = 0.22$ ) explained the independently extracted neural signal (paired t-test:  $t_{22} = 2.52$ ,  $P = 0.02$ , Figure 5.3A). Furthermore, relationships between items arranged in a map-like structure are non-directional. Our subjects were not constrained to experience each pair of transitions an equal number of times (Figure 5.3B). Based upon this we could test whether the fMRI signal was better predicted by the true or symmetrized distance between any two objects. We constructed a measure of the shortest path between each pair of objects according to the actual number of times each transition was experienced by a subject during training (Online Methods). When allowing this measure to compete with its symmetrized self in a linear model, it was the symmetrized version alone that predicted the fMRI suppression effect (Figure 5.3C,  $t_{22} = 2.66$ ,  $P = 0.01$  and  $t_{22} = -1.61$ ,  $P = 0.12$ ).



**Figure 5.3 Relational information is organized as a map.** **A** Linear regression on neural activity with number of links and average time between two objects during training as regressors,  $t_{22} = 2.52$ ,  $P = 0.02$ . **B** Absolute difference in the number of times a transition was visited in one vs. the other direction (e.g. 5 preceded by 1 vs. 1 preceded by 5) for all subjects. **C** Multiple linear regression on neural activity with the shortest path between objects, and the symmetrized shortest path between objects as regressors. **D** 7x7 matrix representing the average fMRI signal in response to an object depending on which other object preceded, averaged across subjects and symmetrized. This matrix was used for the MDS visualized in **E**. **E** Visualization of the localization of the object representations in a 2-dimensional space according to multiple-dimensional scaling. Lines indicate transitions experienced during training. All analyses were performed on data extracted from the peak MNI coordinate reported in Chadwick et al. (2015),  $[-20, -25, -24]$ . Error bars show mean and s.e.m.; a.u., arbitrary units.

In order to test whether these map-like features are a consequence of a map-like organisation, we organised the signal into a 7x7 matrix with each matrix element reflecting the mean fMRI response across subjects to transitions between the corresponding pairs of objects (Figure 5.3D). For example, element [2,7] in this matrix is the average response across all subjects when they see object 7 on the graph preceded by object 2. Because the signal is suppressed for nearby objects, this matrix is analogous to a distance matrix. When we applied multidimensional scaling to visualise the most faithful 2-dimensional representation of distances in this matrix, the graph structure of our experimental map was recovered despite the subjects' professed ignorance of any such organisation (Figure 5.3E). Notably, the data were extracted from an independent ROI taken from an experiment investigating maps in allocentric physical space (Chadwick et al., 2015).

## 5.5 Discussion

The hippocampal-entorhinal system is engaged when an animal navigates in a physical environment and acquires flexible knowledge about spatial relationships. In mammals, the hippocampal-entorhinal system contributes to spatial navigation by mapping spatial relationships in situations where knowledge is physical, continuous and consciously available (Chadwick et al., 2015; Derdikman and Moser, 2010; Howard et al., 2014; Spiers and Maguire, 2007). Here, we use a statistical learning paradigm to demonstrate the entorhinal cortex also efficiently extracts statistical regularities in a non-spatial task where the relationships between items are discrete, and organizes this non-spatial relational knowledge in an abstract relational map, suggesting the hippocampal-entorhinal system creates metric representations of discrete relationships that are completely unlike relationships in physical space (Eichenbaum and Cohen, 2014). These results add to the notion that the hippocampal formation maps experiences across a wide range of different dimensions, thereby supporting flexible behaviour across many domains of life (Schiller et al., 2015).

Using models of the environmental structure to enable transitive inferences is an essential component of intelligent reasoning across species and cognitive domains (Tervo et al., 2016). For example, rats use spatial knowledge to rapidly integrate new spatial information (Tse et al., 2007), songbirds can detect the violation of artificial grammar rules (Abe and Watanabe, 2011), and many species can infer social relationships between conspecifics (Bond et al., 2003; Grosenick et al., 2007; Kumaran et al., 2012; Paz-Y-Miño C et al., 2004). Here, we

demonstrate that this cognitive flexibility is underpinned by a metric organization of the relationships between non-spatial stimuli. In physical space, the precise metric between landmarks provided by entorhinal grid cells enables the rapid and flexible computation of distances and novel paths through the environment (Bush et al., 2015; Stemmler et al., 2015). Our results suggest that the navigation through an abstract concept space between discrete experiences benefits from a very similar mapping of relationships, even in situations where these are discrete, non-spatial and unavailable to self-reported awareness. A map of abstract relationships, where distances between associative states are accurately reflected independently from their behavioural relevance, is very useful in situations where learned relationships can inform novel inference.

It is worth noting that in situations where a map is used for guiding behaviours that maximize future rewards, planning can be facilitated by additionally encoding the task-relevant aspects of states or transitions in the map. For example, if a state has previously been paired with a reward, it can be beneficial to reflect in the map itself that this state and all the transitions leading to it are more valuable than other, equidistant transitions. Encoding this information directly in the neural representation of the world structure would be computationally efficient, as the value of a state or a transition does not need to be computed explicitly when the map is later used for planning a trajectory. Indeed, in the reinforcement learning literature, it has been proposed that state representations are inherently prospective, such that states that make similar predictions about future rewards have a similar neural representation (e.g. ‘successor representation’, Dayan, 1993). Such an account would result in a very different kind of distance metric from the one observed here, because states that make similar predictions about the future are representationally more similar than equidistant states of lesser value (Stachenfeld et al., 2014).

In physical space, it is well established that experience and behavioural relevance influence place cell firing. For example, place cells encode prospective information about the animal’s trajectory in the immediate future (Ainge et al., 2007; Ferbinteanu and Shapiro, 2003) and differentiate between identical paths in space if the required action at the end of the path differs (Wood et al., 2000). Furthermore, the concentration of place cells is particularly high around reward locations (Hok et al., 2007; Hollup et al., 2001), a phenomenon which aids memory recall (Dupret et al., 2010). The organization of the spatial cognitive map with respect to behavioural relevance is presumably mediated via the input the hippocampus receives from a wide range of other brain areas that are active during learning, including input

from sensory cortices and dopamine projections from the midbrain. For example, value information is propagated from a rewarded location backwards along an experienced trajectory by simultaneous reverse replay of spatial sequential activity patterns and a VTA prediction error signal during non-exploratory wake periods immediately after spatial experience (Gomperts et al., 2015). It is conceivable that similar mechanisms are at play for abstract relational maps, resulting in a direct encoding of behavioural relevance within the map representation. It remains unclear whether a warped map representing behavioural relevance as well as map representing associative distance co-exist in the brain and interact to guide decision making.

It also remains unclear how exactly the relational map aids goal-directed behaviour. One intriguing hypothesis is one whereby the hippocampal-entorhinal system stores an abstract cognitive map of the world, and orbitofrontal cortex (OFC) represents an animal's current location within this space and guides goal-directed decision making (Wilson et al., 2014). OFC is critical for encoding stimulus-reward associations (Klein-Flügge et al., 2013a) and for credit assignment, whereby an outcome is specifically attributed to the relevant choice (Walton et al., 2010). If an association between stimuli and rewards has to be updated flexibly, OFC activity modulates the strength of an association between stimuli and outcomes that is stored in the hippocampus (Boorman et al., 2016). OFC is thus critical for identifying the currently relevant state of the world, associating it with reward and using this knowledge to guide behaviour. It is particularly well suited for assigning a stimulus or an action to an outcome through its access to multisensory and emotional information, memories and rewards. Direct access to a model representation stored in the hippocampus may be driven by the strong anatomical connections between the two anatomical structures (Carmichael and Price, 1995).

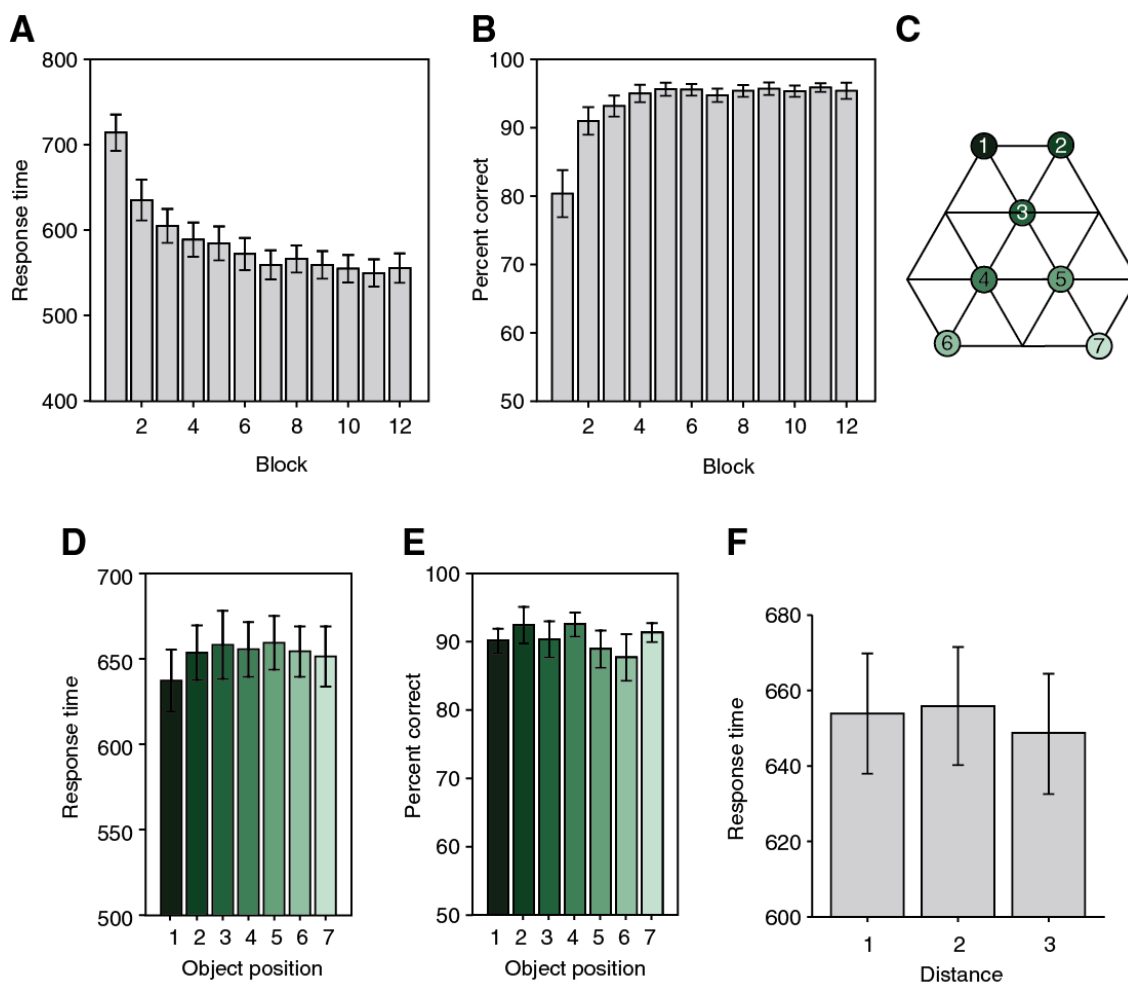
In our experiment, we found no evidence for a representation of the relational structure in prefrontal cortex. This may be related to the fact that training in our task was implicit and occurred while participants were performing an independent cover task. Subjects were not made aware of structure in the object sequence and the structure was irrelevant for behaviour. This setting is reminiscent of the *latent learning* experiments performed by Tolman in the 20<sup>th</sup> century (Tolman and Honzik, 1930), where animals constructed a 'cognitive map' of the environment even in the absence of reinforcement. This map can then be used to enable rapid goal-directed learning when a reward is later introduced (Tolman, 1948). Future research will be needed to investigate how relational information mapped in the

hippocampal-entorhinal system is manipulated in prefrontal cortex in situations where the map is useful for guiding goal-directed behaviour.

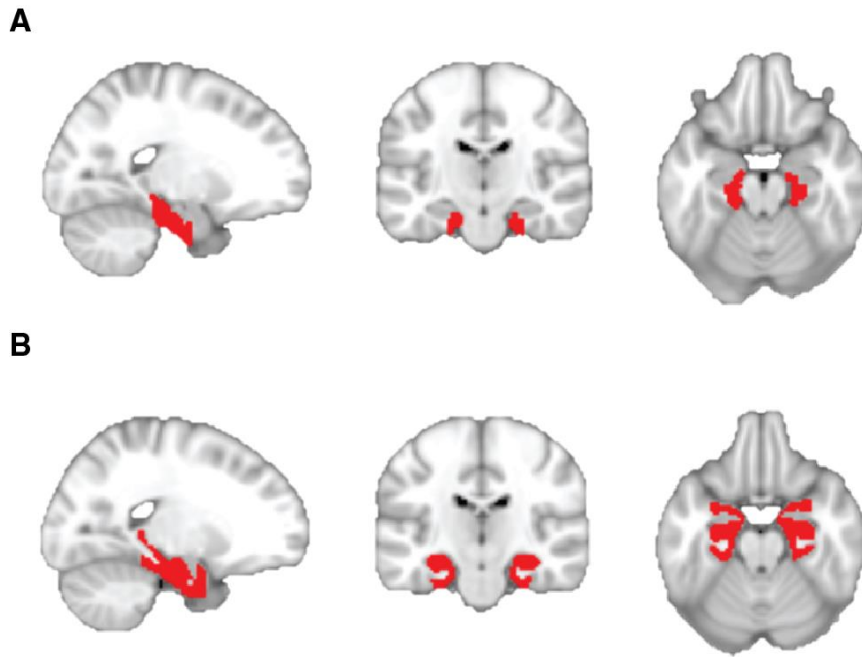
We specifically find a map-like representation in the entorhinal cortex. The most abundant cell type in medial entorhinal cortex are grid cells, characterized by very regularly spaced triangular firing fields (Hafting et al., 2005). In physical space, grid cells represent a context-independent spatial metric for path integration and vector navigation (Bush et al., 2014). Our findings suggest that grid cells can also represent distances or vectors between non-spatial and discrete, arbitrary representational states. A recent theoretical analysis (Dordek et al., 2016) and my own simulations in Chapter 6 suggest that this distance-dependent modulation of representational similarity may be the consequence of grid cells encoding the covariance structure of an environment, in a situation where the hippocampus proper encodes individual experiences. However, it is also conceivable that a plasticity induced between neighbouring stimuli on the graph due to their repeated temporal co-occurrence results in a decrease in representational similarity with distance (Chapter 6). Due to the stimulus randomization during the scanning procedure the only difference between trials was the association of a stimulus with the preceding stimulus during training. The distance effect we observe can therefore not result from visual differences or pre-existing associations between objects.

Maps are well established in the entorhinal cortex in humans and rodents for spatial situations where knowledge is physical, continuous and consciously available (Chadwick et al., 2015; Howard et al., 2014). Here, we demonstrate the entorhinal cortex contains complex maps in situations where knowledge is abstract, discrete and not accessible to self-reported awareness. Such an organization of relational information might be the basis for an animal's ability to navigate through an abstract concept space and perform flexible computations without direct experience.

## 5.6 Supplementary Figures

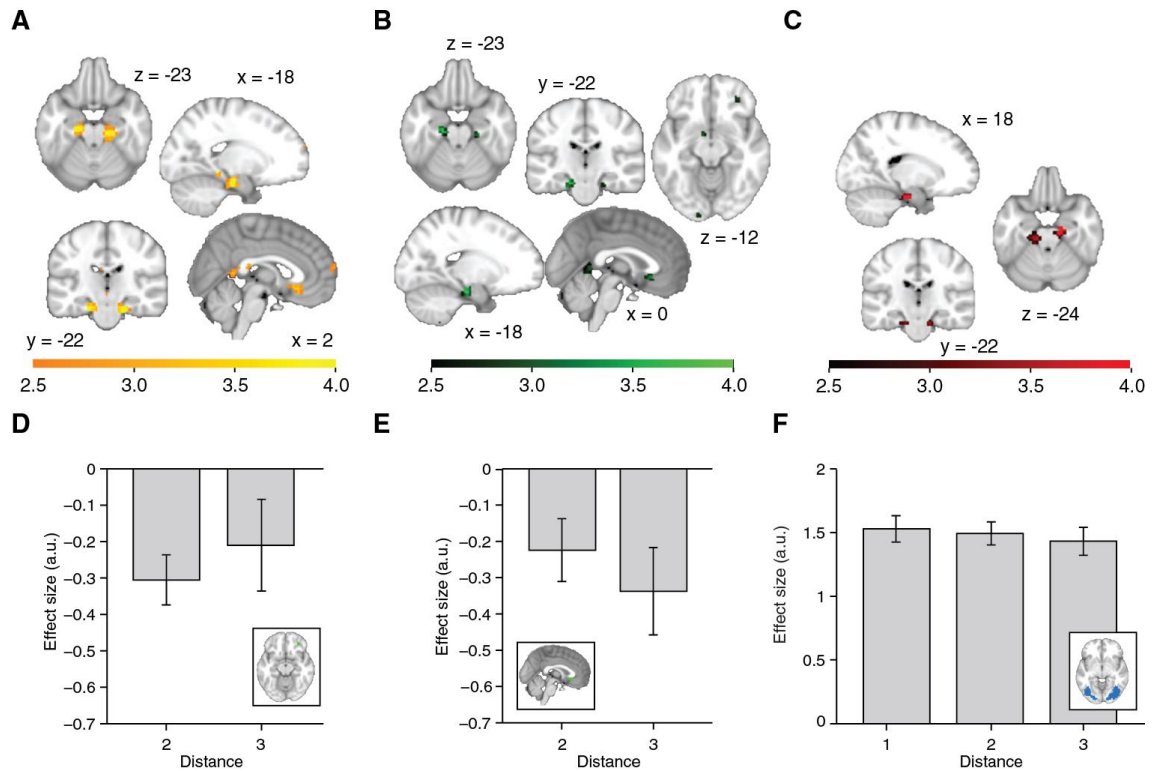


**Supplementary Figure 5.1 Task performance.** **A** Reaction times and **B** performance on the orientation judgment cover task performed during training on day 1 for each of the twelve blocks. **C** Graph structure indicating the object position. **D** Response time ( $F_{6,132} = 0.68$ ,  $p = 0.67$ ) and **E** performance ( $F_{6,132} = 0.56$ ,  $p = 0.77$ ) on the patch detection cover task performed during the fMRI experiment does not differ for the different object locations. **F** Response times on the patch detection cover task does not depend on the distance between objects on the graph ( $F_{2,44} = 0.46$ ,  $p = 0.63$ ). Error bars show mean and s.e.m.; a.u., arbitrary units

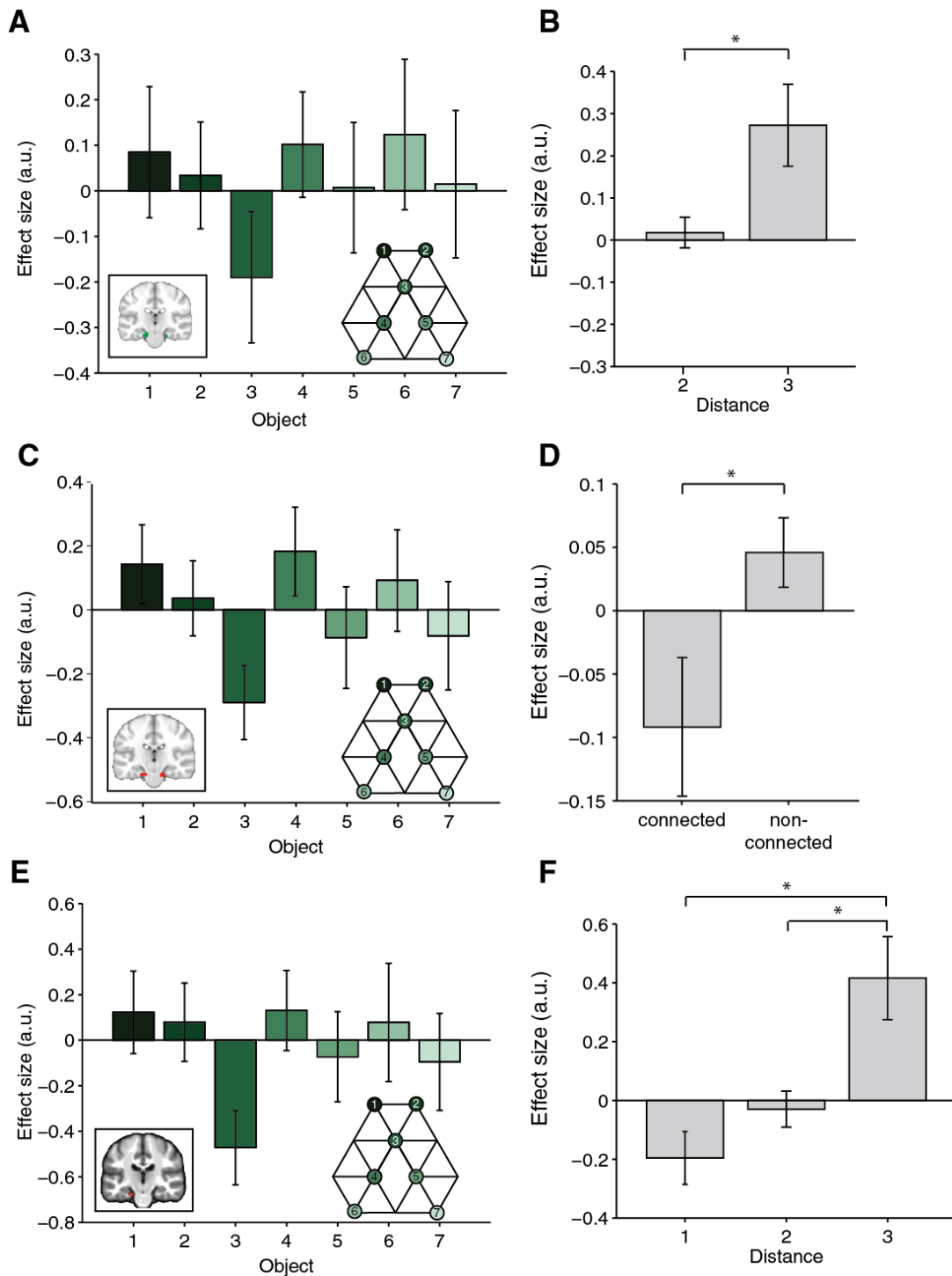


**Supplementary Figure 5.2 Anatomically defined regions of interest used for small-volume correction.** **A** Mask comprising bilateral entorhinal cortex and subiculum, received with thanks from Martin Chadwick (Chadwick et al., 2015). **B** Mask comprising bilateral entorhinal cortex, hippocampus and parahippocampal cortex. Regions were defined using the maximum probability tissue labels provided by Neuromorphometrics, Inc (<http://Neuromorphometrics.com>)





**Supplementary Figure 5.3 Distance-dependent scaling of neural activity is specific to the entorhinal cortex.** **A** All areas displaying a decrease in fMRI adaptation with graph distance. A cluster in subgenual cortex did not survive whole-brain correction for multiple comparisons ( $p_{\text{unc}} = 0.0008$ ,  $p_{\text{FWE}} = 0.99$ , peak  $t_{22} = 3.58$  [3, 23, -10]). **B** All areas displaying suppression for connected stimuli relative to non-connected stimuli on the graph. In addition to the entorhinal clusters reported in the main text, a cluster in orbitofrontal cortex (OFC, peak  $t_{22} = 2.93$ , [30, 41, -10]) and subgenual cortex (peak  $t_{22} = 3.33$ , [0, 23, -7]) were used to define ROIs to test distance-dependent scaling, see **D,E**. **C** All areas displaying greater suppression if a preceding stimulus was two links away than three links away. No areas outside the hippocampal-entorhinal system showed this effect. **D** Parameter estimates for link 2 vs link 3 transitions extracted from the orbitofrontal cortex ROI in **B**. The difference is not significant ( $t_{22} = 0.84$ ,  $p = 0.41$ ). **E** Parameter estimates for link 2 vs link 3 transitions extracted from the subgenual cortex ROI in **B**. The difference is not significant ( $t_{22} = 1.29$ ,  $p = 0.21$ ). **F** Activity in visual areas does not change with distance between items on the graph ( $F_{2,44} = 0.74$ ,  $p = 0.49$ ). Parameter estimates in **F** were extracted from an ROI defined from a contrast indexing a main effect to any visual event in all three blocks (see inset), averaged across subjects. **A-C** are thresholded at  $p < 0.01$  for visualization. Error bars show mean and s.e.m.; a.u., arbitrary units.



**Supplementary Figure 5.4 The distance-dependent scaling cannot be driven by a main effect of object position.** Position-specific activity in the ROI defined according to **A** a connected < non-connected contrast (ROI 1, Figure 5.2B, green,  $F_{6,132} = 0.88$ ,  $p = 0.5$ ), **B** a link 2 < link 3 contrast (ROI 2, Figure 5.2B, red,  $F_{6,132} = 1.96$ ,  $p = 0.08$ ) and **E** the peak coordinate in Chadwick et al. (2015), (ROI 3,  $F_{6,132} = 1.9$ ,  $p = 0.09$ ). The tests reported in Figure 5.2C-E are also significant if performed after removing object-specific activity from the neural data. This was achieved by subtracting the mean activity for each object before testing for **B** a difference in activity for link 2 vs. link 3 transitions ( $t_{22} = 2.10$ ,  $p = 0.048$ ), **D** a difference in activity for connected vs. non-connected stimuli in ROI 2 ( $t_{22} = 2.39$ ,  $p = 0.03$ ) or **F** a distance effect in ROI 3 ( $F_{2,44} = 6.68$ ,  $p = 0.003$ ). Error bars show mean and s.e.m.; a.u., arbitrary units

## **6 MODELS OF MAP-FORMATION IN THE HIPPOCAMPAL- ENTORHINAL SYSTEM**

## 6.1 Abstract

Goal-directed behaviour requires a neural representation of the associations between objects, events and other types of information. In Chapter 5, I demonstrate that such non-spatial and discrete relational information is encoded as a map in human entorhinal cortex. However, the mechanism underlying the formation of such a map-like representation for discrete information remains unclear. Here, I propose two mechanisms that can account for a decrease in representational similarity with associative distance. Firstly, a map representing information about the distance between elements may emerge in a simple Hopfield network with auto-associative attractors and local Hebbian plasticity between pairs of associated objects. Secondly, I propose a framework, whereby grid cell firing patterns and a distance-dependent scaling of representational similarity can result directly from a simple eigendecomposition of place cell activity. While both mechanisms separately account for the distance-dependent scaling of neural similarity, they are not mutually exclusive and may act in concert to encode complex associative structures in the brain.

## 6.2 Introduction

The hippocampal-entorhinal system encodes a cognitive map of space that is used in spatial navigation (O'Keefe and Nadel, 1978). The neural substrate of a spatial cognitive map includes specialized cell types such as hippocampal 'place cells', whose firing is precisely localized in space (Ekstrom et al., 2003; O'Keefe and Nadel, 1978) and entorhinal 'grid cells' with hexagonally arranged firing fields (Hafting et al., 2005; Jacobs et al., 2013). These neural codes are well established in rodents and humans for spatial situations where knowledge is physical, continuous and consciously available (Chadwick et al., 2015; Howard et al., 2014; Spiers and Maguire, 2007). In Chapter 5 I demonstrate a map-like organisation can also be extracted from fMRI responses in entorhinal cortex for relationships which are non-spatial rather than spatial, discrete rather than continuous and unavailable to self-reported awareness. This suggests that the hippocampal-entorhinal system also implicitly organizes abstract relational knowledge in a map. Such a representational structure facilitates the computation of relationships between items that have never been directly experienced together, thereby enabling flexible behaviours such as novel inference. However, the mechanisms underlying the formation of a map for discrete relationships remain unclear.

Here, I propose two mechanisms potentially underlying the formation of a map-like organization of abstract relational knowledge. Firstly, the map could be the consequence of pairwise plasticity between the representations of objects that are neighbours on the graph. It is well established that the neural mechanisms underlying the association of pairs of stimuli involve an increase in similarity of the respective stimulus representations. For example, in the human medial temporal lobes (MTL) the similarity of stimulus representations increases for stimuli that were frequently seen in close temporal succession relative to stimuli that rarely co-occurred (Schapiro et al., 2012). Similarly, after repeated exposure to the same stimulus sequence, neurons in macaque anterior ventral temporal lobes respond similarly to stimuli that were neighbours in the sequence (Miyashita, 1988). This suggests that the temporal co-occurrence of stimuli results in the formation of an association, or an increase in similarity of respective neural representations. Such an account is consistent with the notion that the hippocampus represents an auto-associative attractor network, where relational information is directly stored in the synaptic weight matrix. Mechanistically, such an association could be achieved by Hebbian plasticity, whereby connectivity between pairs of neurons forming the respective stimulus representations is modulated in an activity-dependent manner (Hebb, 1949) through LTP (Bliss and Lomo, 1973). Hippocampal CA3 contains strong and largely random, excitatory recurrent connections from other CA3 neurons. In combination with a very high sensitivity to long-term potentiation (LTP), the divergent and convergent loops in CA3 make this substructure ideal for forming auto-associative memories, and for recovering previously stored patterns from partial cues. It is conceivable that global knowledge about the relationship between non-associated objects emerges as a global consequence of pairwise plasticity between temporally co-occurring stimuli.

One of the first models for the associative nature of memory are Hopfield network models (Hopfield, 1982). Hopfield networks are recurrent neural networks which can act as auto-associative attractors and thereby recover a stored memory pattern from a corrupted memory cue. The attractors are created by updating synaptic weights between pairs of model neurons according to the same Hebbian learning rule that is likely to underlie the formation of associations in the hippocampus (Hebb, 1949). However, while traditional Hopfield networks can successfully recover individual memory representations, they fail to account for the correlation between associated stimuli. Because of the auto-associative nature of the network, the states to which the network converges are uncorrelated if memories stored in this network are uncorrelated. Nevertheless, associations can be modelled by adding a

pairwise plasticity term between memory patterns to the auto-associative term (Griniasty et al., 1993). This modulation of a typical Hopfield network model has successfully reproduced the increased representational similarity for neighbouring stimuli in a sequence of visual stimuli (Miyashita, 1988). Here, I investigated the consequences of such pair-wise plasticity for the representation and the emergence of global knowledge about a complex structure formed by multiple association pairs.

While this manipulation successfully reproduces the distance-dependent scaling of representational similarity, a second mechanism could underpin the map-like formation of relational information in the entorhinal cortex reported in Chapter 5. It has traditionally been thought that grid cells reside upstream from hippocampal place cells, and a linear combination of multiple grid cells with various spatial frequencies and random phases could underlie the formation of place fields in the hippocampus (Fuhs and Touretzky, 2006; McNaughton et al., 2006; O'Keefe and Burgess, 2005). However, recent evidence from hippocampal lesion studies (Bonnievie et al., 2013) as well as the observation that place cells develop before grid cells (Langston et al., 2010; Wills et al., 2010) suggest that feedback projections from hippocampal place cells might contribute to the formation of grid cells. Here, I propose that grid cells may perform an eigendecomposition of place cell activity and thereby represent a compressed representation of the environment (Dordek et al., 2016; Stachenfeld et al., 2014), which can be used for guiding goal-directed behaviour. Strikingly, the eigenvectors resulting from an eigendecomposition of simulated place cell activity during a trial sequence in which discrete locations on a graph are visited in random order resembles the hexagonal firing fields observed in entorhinal grid cells during navigation in physical space. Furthermore, the activity pattern of these model grid cells across multiple spatial scales correlates with distance measures, thereby providing a potential explanation for the distance-dependent scaling observed in the entorhinal cortex as described in Chapter 5.

While the two mechanisms we propose are not directly related to one another, it is possible that variants of both accounts exist in parallel in the hippocampal formation, each contributing to the encoding of a complex associative structure.

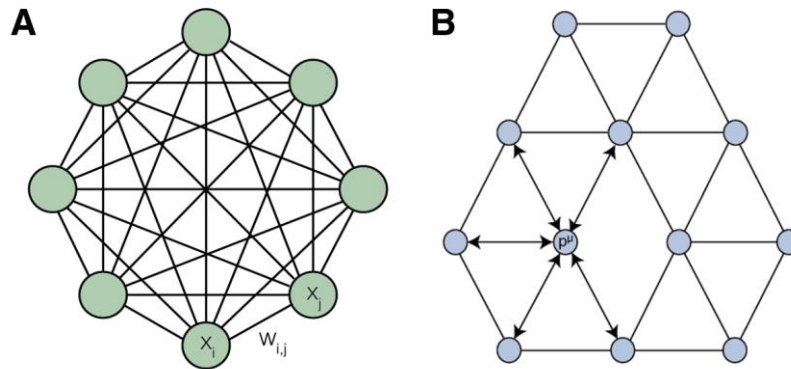
## 6.3 Methods

### 6.3.1 Hopfield network with pairwise plasticity

I set up a fully connected Hopfield network consisting of 6400 neurons (Figure 6.1A) and generated twelve random pattern vectors  $\vec{x}^\mu$ , where  $\mu = 1, 2, \dots, 12$ , with  $-1 \leq x_i \leq 1$ . The (symmetric) weight matrix  $W$ , denoting the connection strength between all pairs of neurons, was defined as:

$$w_{i,j} = \frac{1}{N} \sum_{\mu} (x_i^{\mu} x_j^{\mu} + a \sum_{\nu} x_i^{\mu} x_j^{\nu}), \text{ with } w_{i,i} = 0 \quad (6.1)$$

The auto-associative term  $x_i^{\mu} x_j^{\mu}$  results in the storage of a memory pattern  $\mu$ . Note that it is accompanied by a plasticity term between a pattern  $\mu$  and all its neighbours on the graph  $\nu$ , namely  $x_i^{\mu} x_j^{\nu}$  (Figure 6.1B), weighed by a parameter  $a$ . This term introduces pairwise association between neighbouring patterns on the graph. In the simulations without pairwise associations (Figure 6.3),  $a$  was set to 0, in all other simulations  $a$  was set to 0.1.



**Figure 6.1 Schematic depiction of the architecture of the complete, undirected Hopfield network.** **A** Network architecture. The connection strength between pairs of neurons  $x_i$  and  $x_j$  is given by the synaptic weight  $w_{i,j}$ . Note the actual network consisted of 6400 neurons rather than 8 as depicted here. **B** Graph structure formed by 12 memory patterns. Plasticity was introduced between neighbouring patterns on the graph, as depicted for pattern  $p^\mu$  and its neighbours. Note this is the same structure used in the fMRI experiment, Chapter 5.

To recall a pattern of activation, the network was initialized with a cue or network state defined as one of the patterns  $\mu$  plus noise, and bounded between -1 and 1, i.e.

$$x_i = \tanh(x_i^{\mu} + N(0,1)) \quad (6.2)$$

The activity of each neuron was then iteratively updated. To this end, the input to each neuron  $x_i$  was computed as the sum of the activity of all other neurons  $x_j$  multiplied by the synaptic weight  $w_{i,j}$  between neurons  $x_i$  and all other neurons  $x_j$ . Again, activity was bounded between -1 and 1. This term was weighed with a factor  $dt$  (here: 0.1) and added to the current state of the neuron  $x_i$ , weighed by  $1 - dt$ :

$$x_i = (1 - dt)x_i + dt * \tanh\left(\sum_j w_{i,j}x_j\right) \quad (6.3)$$

For each pattern we performed 20 rounds of simulations, in which the network was initialized with a novel cue. The network's state was assessed at each iteration, and the simulation ended after 500 iterations.

To probe memory recollection, we correlated the network states across iterations with the 12 original memory patterns. To investigate distance effects in the network, a neural similarity measure was defined between the network states initialized with different memory cues. This similarity measure corresponded to the average correlation between network states at each iteration, averaged across the 500 iteration steps. The resulting neural similarity measures between pairs of memory cues could subsequently be sorted according to the distance between any two memories on the graph.

To extract the Euclidian distance between objects on the graph multi-dimensional scaling (MDS) was performed on a matrix denoting the number of links between all pairs of memories. MDS arranges objects spatially such that the distances between them in space approximate their similarities as defined by a distance matrix. From the resulting spatial arrangement, a Euclidian measure could be extracted, which was used for the correlation analysis in Figure 6.5. Crucially, the Euclidian distance measure varied within groups of link distances, i.e. the Euclidian distance resulting from MDS is not the same for all memories that are 1 link away from each other. Therefore, we could correlate the Euclidian distance measure between memories belonging to the same group of link distances (all memories that are 1 link, 2 links, 3 links or 4 links away from each other) with the neural similarity measure between the corresponding network states. To more directly test whether shortest path or Euclidian distance explain the structure in the neural similarity measures between network states, a multiple linear regression was performed with the number of links and the Euclidian



distance between two objects as regressors. These analyses were performed separately for each of the 20 simulations and averaged across simulations.

Furthermore, we performed MDS directly on the neural similarity measure, averaged across 20 simulations, to visualize the relationship between network states in an abstract memory space. More specifically, we performed MDS on a matrix denoting 1 minus the mean correlation between mean network states across iterations and simulations for each memory pair. For example, element 2-5 in the matrix corresponded to 1 minus the average correlation between the average network states for patterns 2 and 5. Because the correlation between network states scales with distance, this matrix effectively corresponds to a distance or similarity matrix. Note that MDS can only be performed on symmetric matrices with positive entries. We therefore normalized the matrix by subtracting the minimum value of the matrix and adding 1.

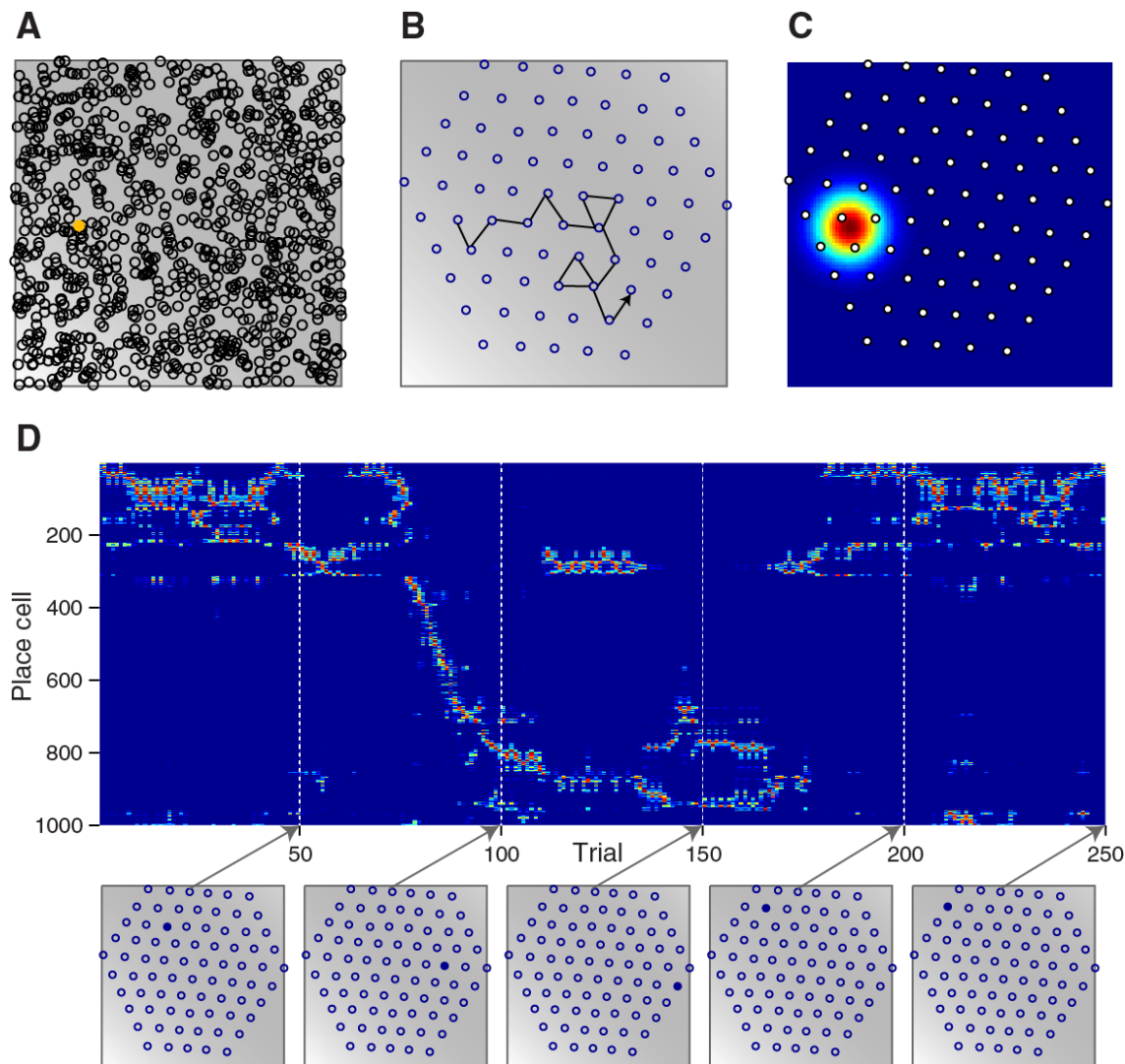
### 6.3.2 Eigendecomposition of place cell activity

In a separate model, 1000 place fields were randomly placed in a quadratic space (Figure 6.2A). Activity of each place cell was simulated based on a multivariate normal distribution with the mean  $\mu$  set to the centre of the place field and covariance  $\sigma$  corresponding to the identity matrix (Figure 6.2C). The graph structure was composed of 75 memories arranged in a hexagonal fashion, whereby each memory had six equidistant neighbours (Figure 6.2B). During a simulated training session, locations corresponding to a node on the graph (Figure 6.2B) were visited in random order, with the only constraint that the next location in a sequence had to be a direct neighbour of the current location on the graph. In total, the sequence consisted of 10,000 trials. On each trial, the activity of each of the 1000 place cells was estimated (see visualized trajectory in Figure 6.2B). Note that each place cell typically fired in response to multiple stimuli due to the size of the place field (Figure 6.2C).

Subsequently, a principal component analysis (PCA) was performed on the cell x trial matrix (Figure 6.2D) and the principal components were projected back into a 2D space. The first 81 components are depicted in Figure 6.7.

A principal component analysis is a linear transformation which projects a high-dimensional dataset onto a set of orthogonal ‘principal components’, sorted according to the amount of variance they explain in the data. In many situations (e.g. if many of the original variables are correlated), the number of principal components needed to explain the variance is smaller than the dimensionality of the original data. This allows for the dimensionality of

the data to be reduced by representing data in terms of their loading onto the principal components.



**Figure 6.2 Simulated place fields.** **A** Peak location of 1000 randomly placed place fields. The yellow dot indicates the location of the place field visualized in **C**. **B** Locations of discrete objects in this space forming the graph. These discrete locations are the only ones visited during the simulation. The black line indicates an exemplary trajectory through this space. **C** Place field of one exemplary simulated place cell. Colour coding indicates simulated firing rate. Note the cell's firing rate decreases with distance from the centre of the place cell. Due to the size of a place field, each place cell responds to multiple locations. Superimposed is the object graph, indicating the discrete object locations in this space in white. **D** Segment of the place cell x trial matrix, indicating each cell's firing rate as different locations are visited during the simulation. Cells are sorted according to their peak firing rate. Locations on the graph are visualized for trials  $t = 50, 100, 150, 200$  and  $250$  (bottom).

Principal components of a matrix  $A$  can be found by performing an eigenvalue decomposition and thereby identifying the eigenvectors and eigenvalues of the matrix. Eigenvectors of a matrix  $A$  are those vectors  $x$  whose direction does not change if they are multiplied by  $A$ , i.e.:

$$Ax = \lambda x \tag{6.4}$$

Because an eigenvector's length is normalized, an eigenvector only describes a direction. However, each eigenvector  $x$  has a corresponding eigenvalue  $\lambda$ , indicating how much variance the eigenvector captures.

One of the advantages of describing a matrix in terms of its eigenvectors and eigenvalues is the fact that  $A^n$  has the same eigenvectors and eigenvalues as  $A$  and can readily be computed as:

$$A^n = x\lambda^n x^{-1} \tag{6.5}$$

In a situation where  $A$  describes an adjacency matrix,  $A^n$  can be understood as a representation of the number of paths of length  $n$  between element  $i$  and element  $j$  of the matrix. The number of paths between two elements, in turn, is a direct correlate of the distance between two elements on a graph: More paths exist between two elements the closer they are located to each other on the graph. Eigenvectors and eigenvalues may therefore be used for rapidly computing possible future states, simply by scaling the weights or eigenvalues (Muller et al., in review).

To compare the pattern similarity for different stimuli on the graph, we set up activity vectors  $[v_1, v_2, v_3, v_4]$  for each location  $v$ , with  $v_i$  corresponding to the coefficient of component  $i$  derived from the PCA at location  $v$ . Two patterns  $v$  and  $z$  can then be compared by computing the angle  $\Theta$  between the two corresponding activity vectors.

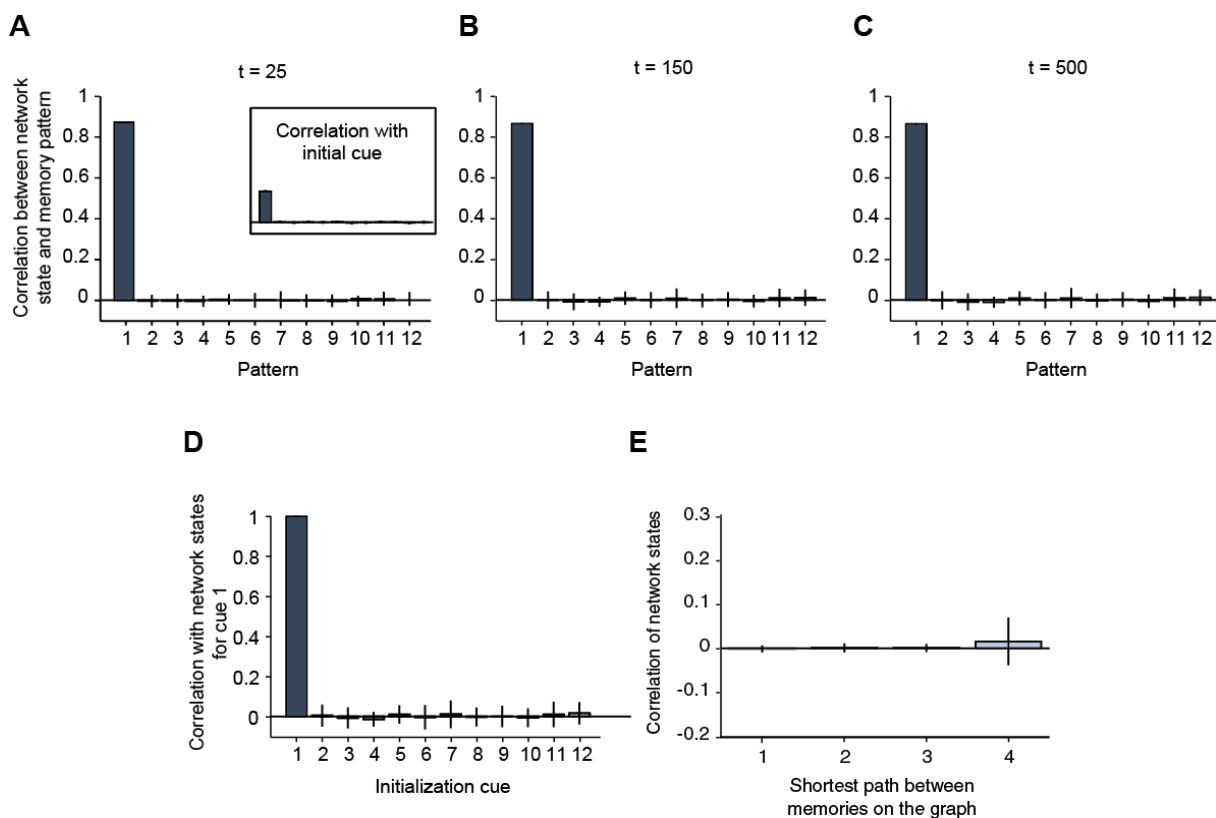
$$\cos \Theta = \frac{v \cdot z}{\|v\| \|z\|} \tag{6.6}$$

where  $\cos \Theta = 1$  if the two vectors are perfectly aligned, and  $\cos \Theta = 0$  if the two vectors are orthogonal. Note this measure directly corresponds to the normalized dot product or correlation coefficient. Angles between stimulus pairs were then sorted according to the distance between the two stimuli on the graph to assess a distance-dependent scaling of representational similarity.

## 6.4 Results

### 6.4.1 A memory pattern can be retrieved from a partial cue in an auto-associative Hopfield network

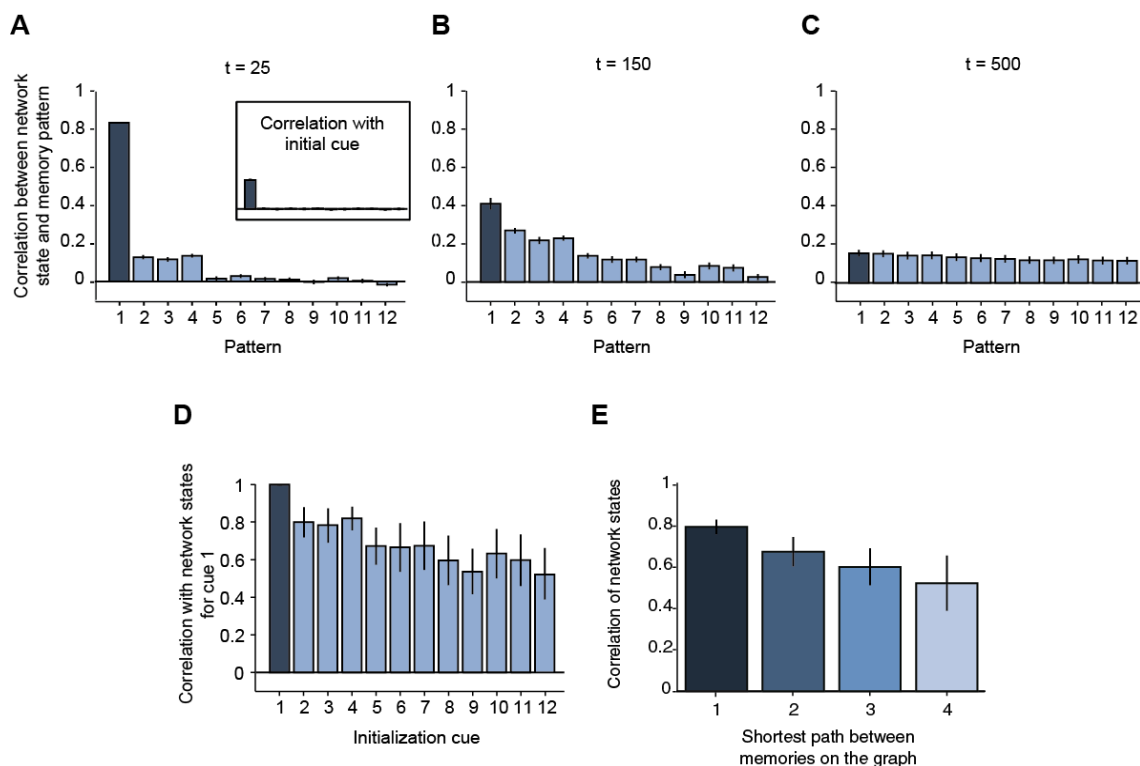
To examine whether global knowledge about a structure can emerge from local plasticity between stimulus-stimulus association pairs, we set up a complete and undirected Hopfield network model (Hopfield, 1982) consisting of 6400 nodes (or “model neurons”, Figure 6.1). I entered 12 memories in the network, by setting all nodes to a specific, but randomly chosen, value. To store these patterns, auto-associative attractors were created by updating the network’s weights according to a Hebbian plasticity rule, where the connection strength between nodes  $x_i$  and  $x_j$  was set to the summed product of the pre- and post-synaptic activity for each memory. The network can then retrieve a memory from partial information, i.e. a cue which resembles one of the stored memories (Figure 6.4A-C). However, since synaptic strength in this basic Hopfield network depends exclusively on memory patterns themselves, uncorrelated stimuli lead to uncorrelated final attractor states (Figure 6.4D) and sorting the correlation between final attractor states according to distance does not reveal a distance dependent decrease in representational similarity (Figure 6.4E).



**Figure 6.3 The network retrieves a memory from a partial cue.** **A** Correlation between network state and memory patterns if the network is initialized with a cue that resembles pattern 1 at time points  $t = 25$  (**A**),  $t = 150$  (**B**) and  $t = 500$  (**C**). Inset in **A** demonstrates that the cue most resembles the stored memory pattern 1 with a mean similarity between the initial cue and memory pattern 1 of  $r = 0.27$ . Note that the network recalls the cued memory and stably encodes it. **D** Correlation between network states when initialized with cues 1-12 and network state when initialized with cue 1, averaged across all time points. **E** Sorting of correlation between correlation of network states according to distance between memories on the graph, averaged across all time points. Results are averaged over 20 independent simulations per pattern. Error bars show the standard deviation.

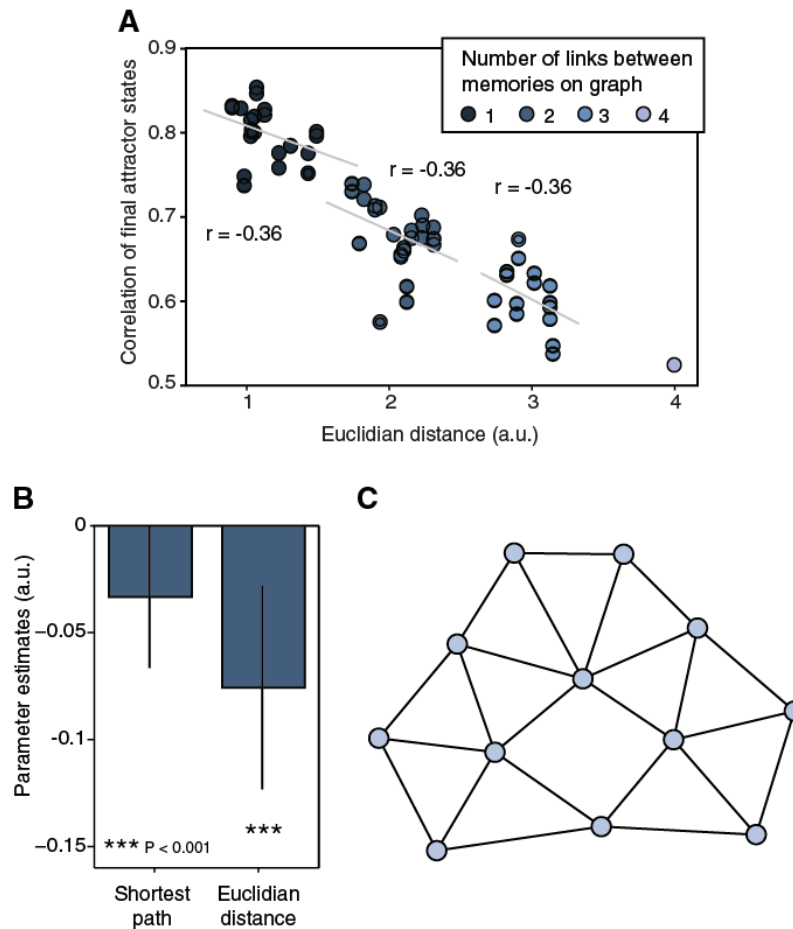
#### 6.4.2 Pairwise plasticity between neighbouring patterns leads to distance-dependent scaling of representational similarity

Neurophysiological studies suggest that a temporal contiguity between neighbouring objects in an object sequence leads to an increase in similarity of underlying neural representations (Miyashita, 1988). To investigate whether such local plasticity between pairs of objects might have consequences for the representation of a structure formed by multiple stimulus-stimulus association pairs, we designed a complex graph composed of multiple pairwise associations between memories (Figure 6.1B). To form stimulus-stimulus association pairs between neighbouring memories on the graph, a plasticity term was added to the synaptic weight matrix for each pair (Griniasty et al., 1993).



**Figure 6.4 Plasticity between neighbouring stimuli on the graph induces correlations between network states.** **A** Correlation between network state and memory patterns if the network is initialized with a cue that most resembles pattern 1 at time points  $t = 25$  (**A**),  $t = 150$  (**B**) and  $t = 500$  (**C**). Inset in **A** demonstrates that the cue most resembles the stored memory pattern 1 with a mean similarity between the initial cue and memory pattern 1 of  $r = 0.27$ . Note that the network initially recalls the cued memory (**A**), but then the correlation with other patterns increases (**B**) and ultimately the network collapses (**C**). **D** Correlation between network states when initialized with cues 1-12 and network state when initialized with cue 1, averaged across all time points. **E** Sorting of correlation between correlation of network states according to distance between memories on the graph, averaged across all time points. The correlation between network states decreases with the shortest path between memories on the graph. Results are averaged over 20 independent simulations per pattern. Error bars show the standard deviation. See also Supplementary Figure 6.1 **Error! Reference source not found.**

While the network state that results from feeding in a particular cue (Figure 6.4A inset) bears greatest similarity to the most similar memory pattern initially (Figure 6.4A), this simple manipulation induces a correlation between the network states for different memory cues, visible after 150 iterations (Figure 6.4B). Ultimately, the network collapses and the network state after feeding in different cues cannot be differentiated (Figure 6.4C, Supplementary Figure 6.1). This is a consequence of an increasing similarity of a network state with all other memory cues (Figure 6.4D). Importantly, across all iterations, the correlation between network states for different cues decreases with the number of links between memories on the graph (Figure 6.4E), suggesting that a global knowledge about the relationship between different memories arises from storing local associations.



**Figure 6.5 Euclidian distances between memories on the graph predict variance over and above the shortest path between memories.** **A** Within each cluster defined by the number of links between memories on the graph, Euclidian distance correlated with the neural similarity of the corresponding network states, averaged across all iterations. **B** Multiple linear regression on correlation between network states. The number of links and the Euclidian distance between memories on the graph were included as regressors competing for variance. Only the Euclidian distance significantly predicted the neural similarity between network states, averaged across all iterations. **C** The network structure can be recovered from the neural similarity measures between network states for different memory cues. Plotted is a two-dimensional representation of 1 minus the correlation matrix, as generated by MDS. Results in **A-C** are averaged over 20 simulations per pattern. Error bars show the standard deviation.; a.u., arbitrary units.

A key feature of a map-like representation of the relationships between landmarks in physical space is the encoding of real-world distances. Similarly, Euclidian distances between discrete memories in an abstract memory space provides a more fine-grained metric of a relationship than the number of links between two memories. To test whether variance in neural similarity is explained by the Euclidian distance between items on the graph over and above the number of links between them, we performed two-dimensional multidimensional scaling on a matrix denoting the shortest path between any two pairs on the graph to estimate the Euclidian distance between two memories on the graph if the distance matrix were to be

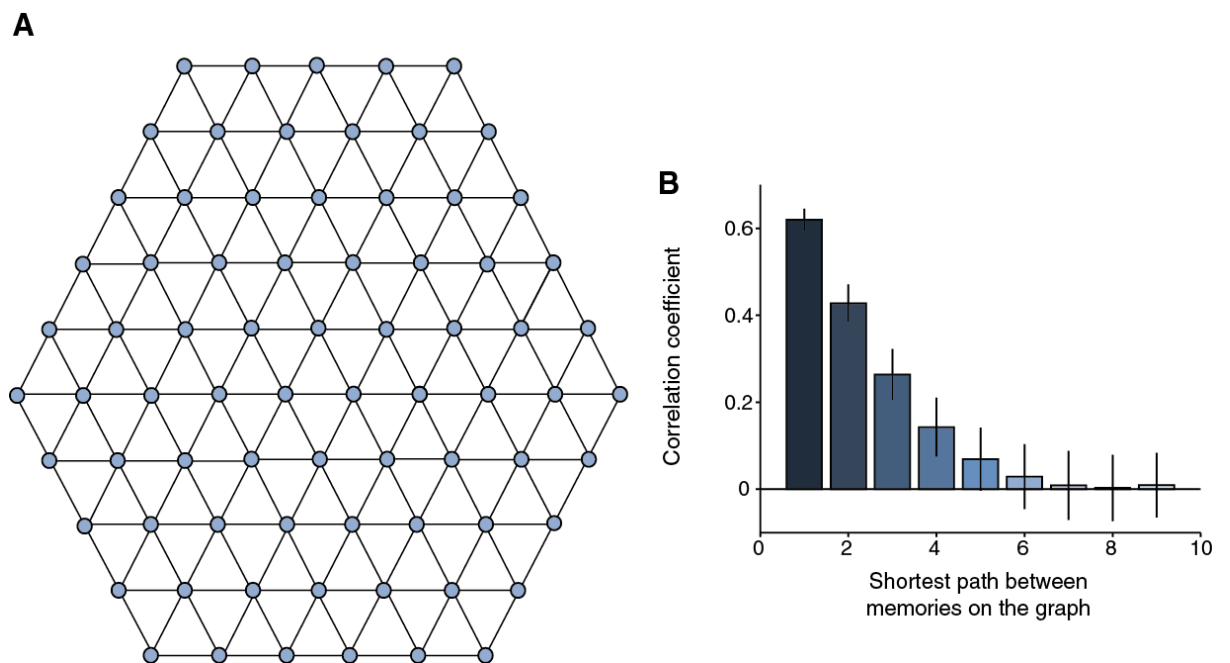
mapped into a 2-dimensional space. I correlated this Euclidian distance measure with the correlation between network states between different cues. Euclidian distance correlated negatively with representational similarity within groups of patterns that are equally far away from each other in terms of number of links, but where the Euclidian distance varied (Dist 1:  $r = -0.40$ ,  $P = 0.006$ , Dist 2:  $r = -0.37$ ,  $P = 0.007$ , Dist 3:  $r = -0.35$ ,  $P = 0.05$ , Figure 6.5A). I confirmed this observation in a multiple linear regression which included the number of links ( $t_{19} = 2.37$ ,  $P = 0.01$ ) and Euclidian distance ( $t_{19} = 5.86$ ,  $P < 0.001$ ) as regressors (Figure 6.5B). This demonstrates that Euclidian distance explains variance over and above the number of links between items, suggesting that simple pairwise plasticity between neighbouring items on a graph results in a representation in which Euclidian distances between items are respected.

In order to test whether these map-like features are a consequence of a map-like organisation, I organised the correlation between network states into a  $7 \times 7$  matrix with each matrix element reflecting the average correlation between pairs of network states across simulations. Because the correlation is higher for nearby objects, this matrix is analogous to an inverse distance matrix. When I applied MDS to visualise the most faithful 2-dimensional representation of this matrix after subtracting it from 1, the graph structure of our experimental map was recovered (Figure 6.5C).

In summary, global knowledge about the discrete relationships between items emerges through increases in representational similarity for pairwise associations in a simple Hopfield network. This knowledge has essential features of a ‘cognitive map’ (O’Keefe and Nadel, 1978), in that Euclidian distances between items are preserved. However, because relationships between memories are encoded within the representation of a memory itself, the representation is unstable and the network ultimately collapses.



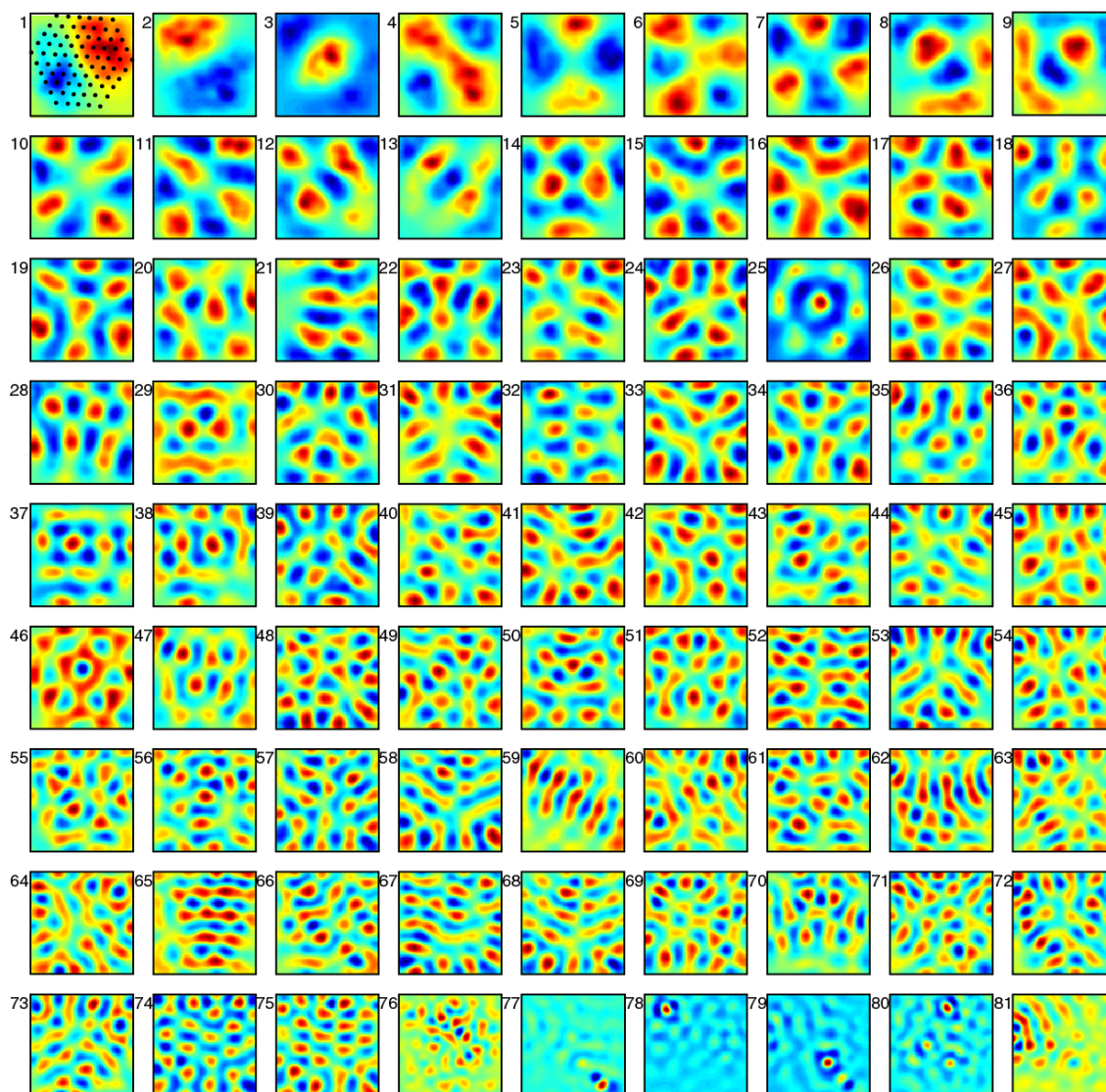
### 6.4.3 An eigendecomposition of the resulting activity patterns reveals grid cell-like activity



**Figure 6.6 Distance-dependent scaling of representational similarity in a large associative graph structure.** **A** Graph structure composed of 75 independent memories. In direct analogy to the smaller graph, pairwise plasticity was introduced between neighbouring items on the graph to model local Hebbian plasticity between neighbouring neural representations. **B** Correlation between network states sorted by distance. In independent runs the network was initialized with cues taken from one of the 75 stored memories. The correlation between the final network states for different cues decreases with the shortest path between memories on the graph. Results are averaged over 20 simulations per pattern. Error bars show the standard deviation.

The same analyses in a much bigger graph composed of 75 independent memories (Figure 6.6A) reveals that the representational similarity between network states decays exponentially with distance (Figure 6.6B). Due to its size, this graph allows for testing an alternative explanation for the distance-dependent scaling of fMRI adaptation observed in Chapter 5. It has been hypothesized that the hexagonal firing pattern of entorhinal grid cells could be the consequence of the computations grid cells perform on place cell inputs. More specifically, it has been suggested that the characteristic six-fold symmetry of entorhinal grid cell firing fields is the consequence of an eigendecomposition of place cell activity (Dordek et al., 2016; Stachenfeld et al., 2014). To investigate what such a spectral decomposition would look like in a situation where the relationships between stimuli are discrete, I simulated the activity of 1000 randomly located place cells, each modelled as 2D Gaussian distributions (Figure 6.2A,C), as the discrete locations on the large graph were visited in a random order (Figure

6.2B). Subsequently, I performed a principal component analysis on the resulting [neuron x trial] matrix.



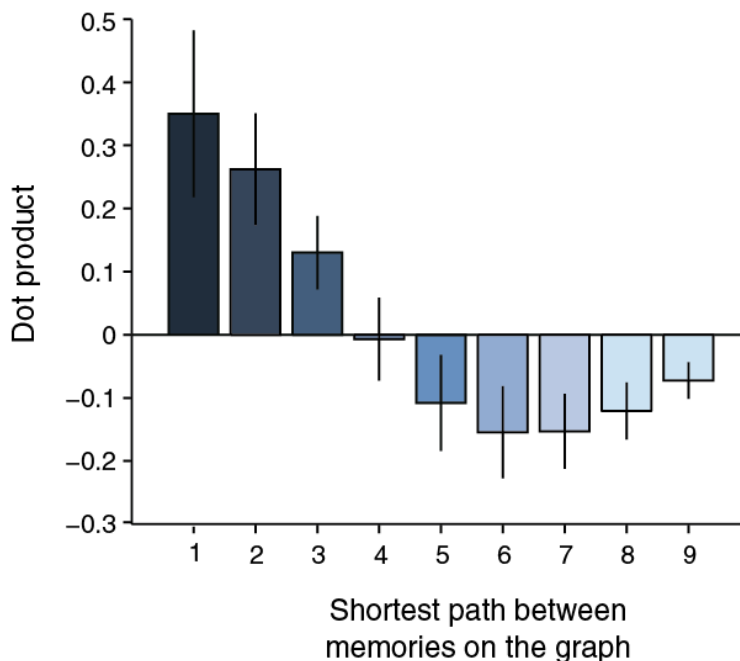
**Figure 6.7 Eigendecomposition of the task structure.** A principal component analysis was performed to extract the covariance structure. The panel shows the first 81 eigenvectors.

To aid an interpretation of the eigenvectors in terms of spatial activity, I projected the PCA eigenvectors back onto the place cell space. While the first few eigenvectors largely divide the space spanned by the graph structure into two or more compartments, some of the high frequency eigenvectors resemble the striking six-fold symmetry observed in entorhinal grid cells (Figure 6.7, e.g. components 19 and 74). Notably, no hexagonal symmetry emerges in situations where the size of the modelled place fields does not span multiple stimuli on the graph (Supplementary Figure 6.2), suggesting that the simultaneous

activation of the same place fields for different stimuli introduces a covariance between neighbouring locations on the graph which is extracted by entorhinal grid cells.

In the brain, grid cells are arranged in discrete modules, each characterized by cells with the same grid scale and grid orientation, but a different grid phase (Stensola et al., 2012). The grid field size increases in discrete steps along the dorsoventral axis (Stensola et al., 2012), corresponding to a posterior-anterior axis in humans (Strange et al., 2014). As a result, grid cells with large fields predominate in ventral MEC, while grid cells with small fields are localized more dorsally. While activity for different locations in space across the entorhinal cortex as a whole is decorrelated due to the orthogonal nature of the principle components, activity for different locations within sub-sections of MEC is likely to be correlated due to this non-uniform distribution of grid cells along the dorsoventral axis. This correlation should be particularly prominent in ventral MEC, where grid fields are large and neighbouring stimuli on the graph lie within the same grid field.

To test whether this results in a correlation between neural patterns that scales with distance between stimuli on the graph for grid cells with low spatial frequencies, I defined a neural activity vector  $[\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4]$  for each location  $\mathbf{v}$  on the graph, where  $\mathbf{v}_i$  corresponds to the coefficient of component  $i$  at location  $\mathbf{v}$ . I then compared activity patterns at different locations by computing  $\cos \Theta$  between the corresponding activity vectors. This measure is directly analogous to a correlation measure, and it approaches 1 the more similar two neural patterns are, and 0 if two patterns are orthogonal. Critically, the size of this similarity measure between pairs of objects decreased with distance on the graph (Figure 6.8), similar to the decrease in representational similarity with distance observed in human entorhinal cortex, Chapter 5. This suggests that the decrease in representational similarity of neural patterns with distance on the graph does not require pairwise plasticity between associated patterns on the graph, but could instead be a direct consequence of an eigendecomposition of place cell activity, which also reveals a striking hexagonal firing field pattern.



**Figure 6.8 Distance-dependence of correlation between activity vectors across the first 4 principal components.** Activity across the first four principal components was computed for each location in space, and the correlation is plotted for pairs of items as a function of distance between them on the graph.

## 6.5 Discussion

Global knowledge about an abstract graph structure formed by associative links is stored in the hippocampal-entorhinal system as a cognitive map (Chapter 5). Here, we present two simple models that can account for map-like representation of this associative structure. According to a simple Hopfield network a cognitive map can emerge from increases in representational similarity for pairwise associations between neighbouring memories on the graph. The model constitutes a simple implementation of a mechanism by which the brain might store the experience-dependent statistical relationships between objects, events or other types of information in our environment directly within the representation of a memory itself. It is also conceivable that the cognitive map encoding distances in physical space might arise from the same principle of pairwise associations between nearby locations in space.

In this model, an increase in representational similarity is achieved by an activity-dependent change in activity, or Hebbian plasticity. The physiological implementation of the Hebbian concept (“what fires together, wires together”), namely LTP, is induced in hippocampal area CA1 by learning (Whitlock et al., 2006), and its maintenance is critical for

storing spatial memories (Pastalkova et al., 2006). However, it is important to note that the model does not address the question how the synaptic structure is learned. Instead, the synaptic matrix is explicitly modified to reflect the correlation between neighbouring patterns on the graph. Various learning mechanisms are conceivable, including a scenario where the network is successively presented with stimuli generated according to a random walk on the graph, in line with the situation the human participants face in the fMRI experiment presented in Chapter 5. Each individual state is then learnt during the time spent in an attractor, and learning of associations occurs by applying Hebbian plasticity to a combination of pre-synaptic activity that represents a preceding pattern and post-synaptic activity that represents a current pattern (Griniasty et al., 1993).

Behaviourally, the change in representational similarity with distance might underlie the phenomenon of priming (Griniasty et al., 1993). Priming constitutes an implicit and nonconscious form of memory where stimulus processing becomes more efficient if a stimulus is repeated, or if a related probe stimulus is presented before a test stimulus (Tulving and Schacter, 1990). Priming has been observed across many cognitive domains, for example in visual (Roediger and McDermott, 1993) or auditory (Schacter and Church, 1992) tasks. It has been hypothesized that it reflects the representational similarity of neuronal attractors, or their proximity in a neuronal state space (Griniasty et al., 1993). Transitions between similar attractors are faster than transitions between attractors that are very dissimilar because fewer neurons need to change their activity profile. It is also consistent with the view that transition times vary with distance on the graph because of a variation of representational similarity with distance. As a consequence, attractors that are easier to reach from a given point are more likely to be visited next in a sequence. This also allows for multiple attractors to be chained, leading to a ‘cell assembly sequence’, or a progression from one cognitive state to the next. Such a cell assembly sequence could be the basis of complex cognitive processes, such as memory recall, planning or decision making and it can result in behavioural patterns that are strongly influenced by the statistical regularities of the environment. Notably, a difference in representational similarity as modelled in this network could also explain the prominent novelty response in the hippocampus (Strange et al., 2005b). The neural representation of a novel stimulus that has never been associated with other stimuli because will be maximally different from other stimulus representations.

However, it is also important to note that the Hopfield network presented here is overly simplified and cannot be considered biologically plausible. Through the pairwise associations

between stored memory patterns, the relationships between different memories are directly encoded within the memory representation itself. This results not only in a corruption of each individual memory, but it also leads to network instability, such that the network collapses after a finite number of iterations. As a consequence, recollection performance for the original memories that are stored in the network decreases as the patterns converge. One reason for the instability of this network is that the plasticity modelled here reflects excitatory connections between neurons alone. In the brain, excitatory connections are precisely balanced by inhibitory connections to stabilize a representation and prevent runaway excitability (Okun and Lampl, 2008). More recently, excitatory-inhibitory Hopfield networks have been developed which reflect these biological constraints and which could be used to more accurately describe the processes in the brain (Amarimber and Amari, 1972). Despite these constraints, the network allows for inferring that a very basic principle of cortical processing, namely activity-dependent plasticity between associated memories, is sufficient to account for a neural signature of a global associative structure. It is also worth noting that the distance-dependence of representational similarity in the network persists throughout all iterations, even if the differences in the network states ultimately become very small.

However, it might be argued that it could be more useful to store memories themselves separately from the relationships between them. To explore this idea further it is worth having a closer look at the relationship between hippocampal place cells and entorhinal grid cells. Whereas hippocampal place cells are characterized by precisely defined firing fields (O'Keefe and Dostrovsky, 1971), entorhinal grid cells are active at multiple spatial locations, because of their hexagonally arranged firing fields (Hafting et al., 2005). Anatomically, grid cells can be found in superficial and deep layers of the entorhinal cortex, and they process inputs to, and outputs from, the hippocampus (Sargolini et al., 2006). Furthermore, the entorhinal cortex contains head direction and border cells, and it is ideally placed to integrate spatial information with inputs it receives from neocortical areas (Hafting et al., 2005). It has therefore traditionally been assumed that MEC neurons project information about spatial location, direction and distance to place cells, with hippocampal place fields constituting a 'read-out' of entorhinal grid cells. Indeed, a linear combination of multiple grid cells with various spatial frequencies and random phases could provide a precise estimate of an animal's location, suggesting that grid cell activity could constitute a basis set that can be combined linearly to generate place fields in the hippocampus (O'Keefe & Burgess 2005, Fuhs & Touretzky 2006, McNaughton et al. 2006). A place field in this setting emerges at the location where most grid cells are in phase.

However, recent evidence is inconsistent with the notion that place cells are downstream read-outs of grid cell activity. Size and shape of place cell firing fields are mostly unaffected by the absence of grid cell inputs (Hales et al., 2014; Koenig et al., 2011) whereas inactivating the hippocampus leads to a loss of hexagonal grid cell firing (Bonnievie et al., 2013). Furthermore, place cells mature before grid cells during development (Langston et al., 2010; Wills et al., 2010) and the development of grid cells coincides with an increased accuracy of place cell activity (Muessig et al., 2015). As a consequence, it has been suggested that grid cells might integrate place field activity (Krupic et al., 2014), for example by performing a spectral decomposition of hippocampal place cell activity (Dordek et al., 2016; Stachenfeld et al., 2014). Such a computation would allow the covariance structure to be extracted from the environment, enabling efficient encoding of relational information. Here, we demonstrate that an eigenvectordecomposition cannot only account for the characteristic hexagonal symmetry of grid cell firing, but it also serves as an explanation for the distance-dependent scaling of neural similarity observed in the fMRI experiment. Note that in this model no pairwise plasticity or other information about associations is introduced explicitly, the covariance structure is a direct consequence of place fields that span multiple locations on the graph. Indeed, if place fields are too small, the eigenvectors display no periodicity. This model is therefore consistent with a notion whereby place cells encode individual episodes, and grid cells encode the covariance structure between those experiences – a model which could apply both in physical space and in an abstract concept space.

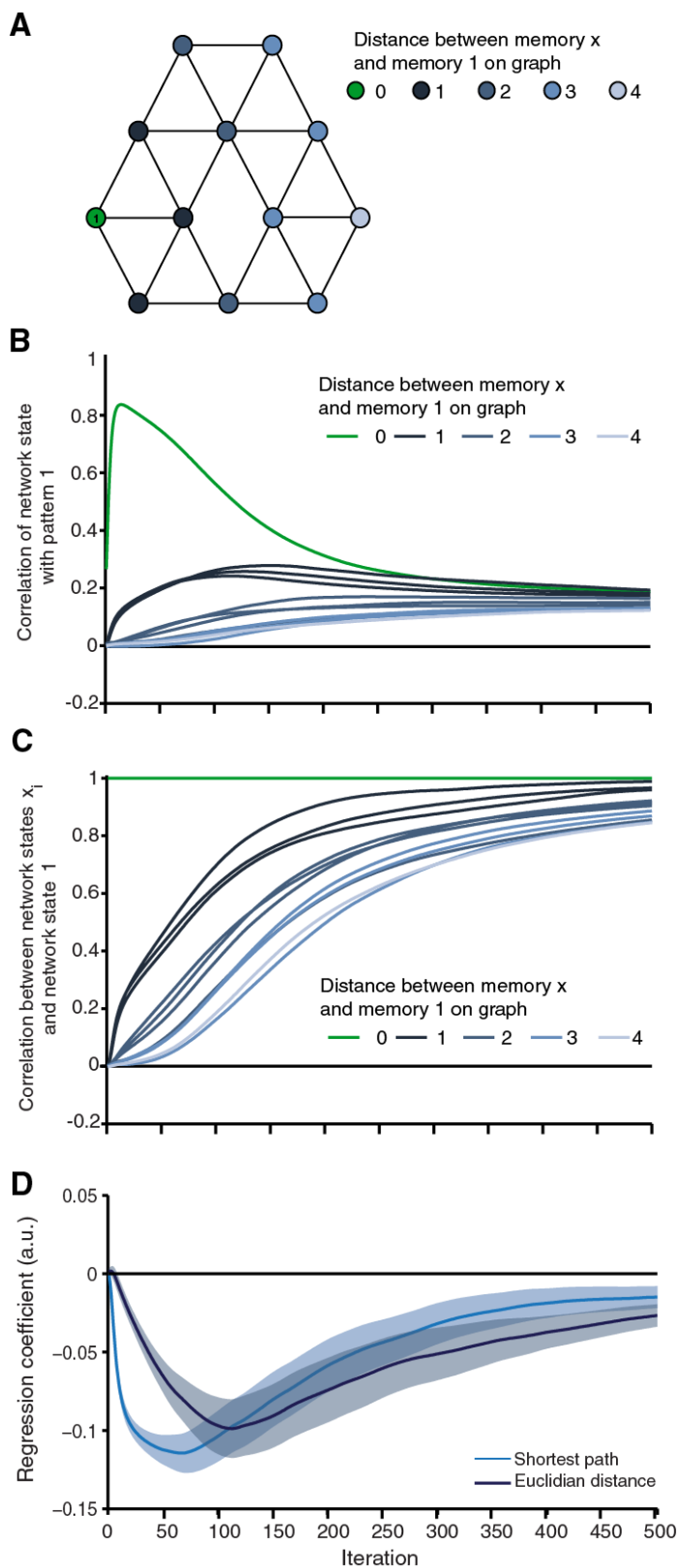
Critically, grid cells are organized in discrete modules, each characterized by a collection of cells whose grid fields are identical in size and orientation, but different in phase (Stensola et al., 2012). The size of grid fields increases in discrete steps along the dorso-ventral axis, with an increase by a factor of 1.4 from one module to the next. Theoretical analyses demonstrate that this organization maximizes the spatial resolution (Mathis et al., 2012). Across grid cells, each location in space is associated with a unique activity pattern, which allows the animal’s precise location in space to be decoded (Moser et al., 2014). However, the modularization along the dorsoventral axis results in a non-uniform distribution of grid cell frequencies that could explain the distance-dependent scaling of activity in areas of the entorhinal cortex with large entorhinal grid field sizes.

In conclusion, we describe two mechanisms which could underlie the emergence of global knowledge about a structure. In a first model, we introduce local pairwise associations which

results in a decrease of representational similarity with distance across the structure as a whole. In a second model, we propose that the grid cell structure observed in the entorhinal cortex can result from an eigenvaluedecomposition of place cell firing. Both accounts can explain the distance-dependent scaling of fMRI adaptation observed in Chapter 5, and future experiments will be needed to understand the precise mechanisms underlying the brain's remarkable ability to extract structure from sensory experiences.

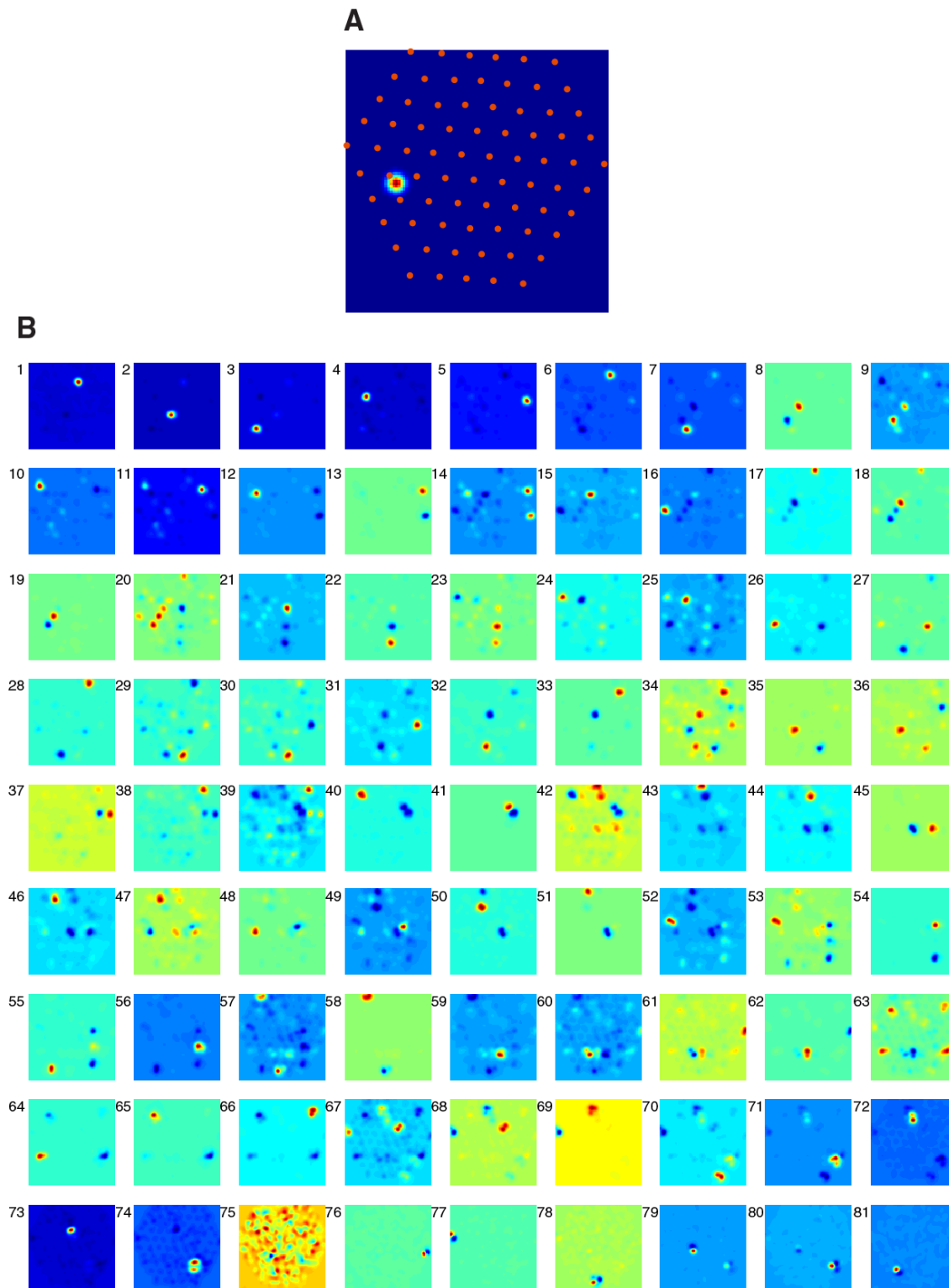


## 6.6 Supplementary Figures



**Supplementary Figure 6.1 Network dynamics.** **A** Graph structure. The colour coding visualizes the distance between any stimulus and stimulus no. 1 coloured in green. **B** Correlation between memory 1 and the network state as a function of the number of iterations, plotted for initializations with cues 1 – 12. The network initially approaches the network state corresponding to memory 1 if initialized with cue 1, but then the similarity between the network state and pattern 1 decreases. During the same time period, the correlation between the network state and pattern 1 increases if the network was initialized with one of the other cues, albeit with a much slower time constant. Ultimately, the correlation between the network state if initialized with cue 1 and the actual memory pattern 1 is no higher than the correlation between memory pattern 1 and the network state if initialized with any other cue, demonstrating that the network collapses after a finite number of iterations. Results are averaged over 20 independent simulations per pattern. Graphs are coloured as a function of the distance between initialization cue and pattern 1, as indicated in **A**. **C** Correlation between the network state if initialized with cue 1 and the network state if initialized with cues 1 – 12 as a function of the number of iterations. All network states become more and more similar to the state the network reaches if initialized with pattern 1. Results are averaged over 20 independent simulations per pattern. Graphs are coloured as a function of the distance between initialization cue and pattern 1, as indicated in **A**. **D** Multiple linear

regression with shortest path and Euclidian distance as regressors as a function of iteration. Despite the fact that the network states become more and more similar, both regressor are still significant at time point 500 ( $t_{19} = 1.78$ ,  $P = 0.046$  and  $t_{19} = 3.44$ ,  $P = 0.001$  respectively) suggesting that the structure remains encoded in the network even if specific memories become blurred.



**Supplementary Figure 6.2 Eigendecomposition of place cell activity for small place fields. A** Place field of one exemplary simulated place cell. Colour coding indicates simulated firing rate. Note the size of the place field is substantially smaller than the size of the place field in Figure 6.2, and each place cell responds to maximally one location on the graph. The graph structure is superimposed. **B** Eigendecomposition of place field activity in response to locations on the graph visited in random order. The panels show the first 81 eigenvectors. Note the triangular lattice structure characteristic for entorhinal grid cells does not emerge in this situation.

## **7 GENERAL DISCUSSION**

## 7.1 Aim

When making decisions it is of critical importance to know which actions to take when. The human brain solves this problem by assigning value to the potential courses of action and choosing the one that leads to the highest expected reward. This strategy requires an adequate model of our environment, which needs to be flexibly adjusted in light of new information. If events are not consistent with the internal model of the world, the model needs to be modified or replaced by a more appropriate one.

This thesis uses representational techniques in combination with computational modelling to understand how such internal models are represented in the brain, and how model updates can influence our behaviour. More specifically, the work investigates how representations that can support goal-directed behaviour are formed and represented in the hippocampal-entorhinal system (Chapters 5-6) and how learning about another person's preferences induces plasticity in value computations and thereby influences decision making (Chapters 3-4). In this final chapter, I conclude by briefly discussing the general implications of these findings for our understanding of information processing in the brain, with a special focus on prediction-error driven learning from experience.

## 7.2 Updating models of the world in social decision making

In Chapter 3, I demonstrated that learning about the preferences of another individual in an intertemporal choice paradigm leads to social influence, whereby subjects own preferences shift towards those of the interaction partner. Control experiments demonstrate this behavioural change in preference is driven by learning about the other and cannot be explained by executing a choice per se. Learning in this context can be modelled by a Bayesian learning algorithm that updates prior expectations about the other's preferences whenever new information is encountered. The Bayesian model can also be used to make predictions of neural activity associated with learning, such as prediction errors participants experience when receiving feedback about the other's choice.

In Chapter 4, I demonstrate that such a prediction error signal is represented in various brain areas, including ventral striatum. I present evidence that this striatal signal drives a mPFC plasticity effect, whereby the neural value representations for self and other become more similar as subjects learn about the other's preferences. The degree of mPFC plasticity

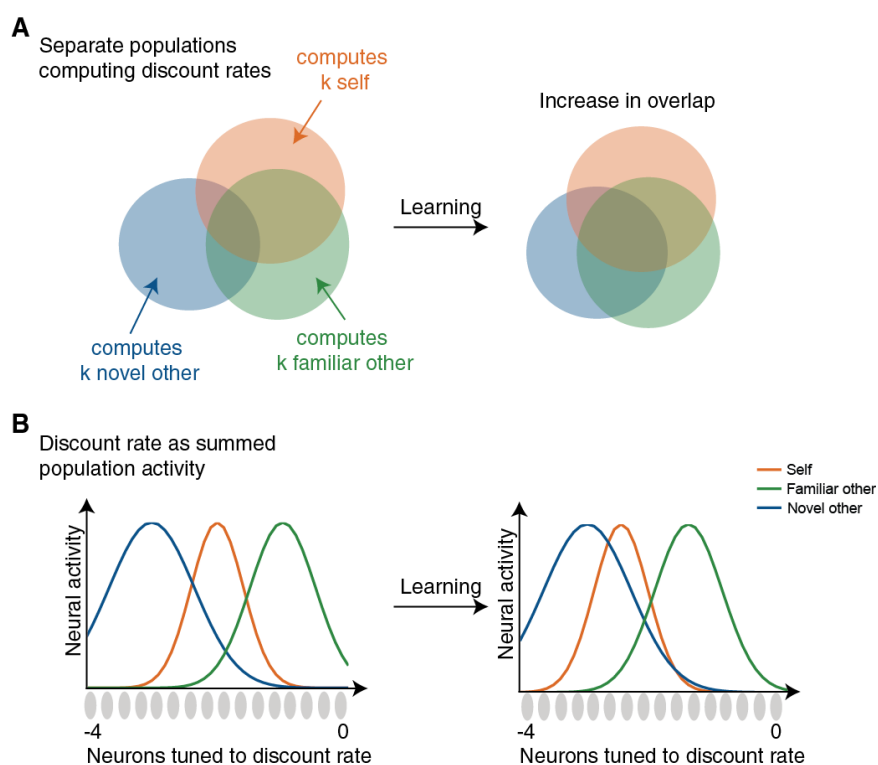
predicts how much subjects' own preferences change, suggesting that inter-individual differences in the malleability of subjective preferences might be the result of a striatal prediction error signal driving mPFC plasticity. This demonstrates that a learning-induced plasticity in a valuation network can underlie social influence in decision making.

Prediction errors in this context can be thought of in two ways. Firstly, prediction errors can arise when the other's true choice is different from the choice a subject predicted the other would make. This 'other-regarding' prediction error can have an important function as a teaching signal for updating a model about the other's values. Secondly, prediction errors can arise when the other's preferences diverge from subjects' own values, i.e. in a situation where the subject would have made a different choice for themselves given the same context. This 'self-regarding' prediction error might be more relevant for influencing subjects' own preferences, and it is indeed this prediction error that is represented in ventral striatum and that drives plasticity in the prefrontal cortex.

A distinction between 'other-regarding' and self-regarding' prediction errors is reported across a range of social neuroscience studies. When subjects rate the attractiveness of faces before and after learning about the opinions of others, striatal activity elicited while observing the others' choices predicts the degree of conformity (Klucharev et al., 2009). When subjects assigned to the role of an 'investor' and a 'trustee' interact repeatedly in a trust game, activity in the trustee's caudate nucleus displays a prediction error-like signal correlating with the amount of money the investor invests. In line with a widely observed shift of a dopaminergic prediction error signal from the time of a reward to the time of the CS, this signal is initially seen at the time when investment is revealed, but occurs before the time the investment has been made in later rounds of the game, suggesting that the trustee has constructed a model of the investor's behaviour (King-Casas et al., 2005). Consistent with the notion of a 'self-regarding' prediction error, however, this signal does not drive learning about the investor's behaviour, but instead signals the trustee's response to the investor's choice. Striatal activity is increased in situations where the trustee returns larger amounts of money relative to situations where he returns smaller amounts of money. Overall, these studies are in line with the observation in this thesis that striatal prediction errors in social contexts signal relevance for updating one's own behavioural policy.

The observation of a plasticity in one's own value representation has implications for our understanding of the organizational structure of mPFC. Our data are not consistent with an account whereby subjects use their own representation in order to simulate other people's

preferences. If subjects had used their own representation to choose on behalf of the novel other throughout the experiment, we would not expect any changes in the overlap between self and the novel other, and therefore no change in repetition suppression between the two agents. If instead a novel representation would be constructed from the representation of self (Barron et al., 2013) we would expect a decrease in repetition suppression between self and novel other over the course of the experiment. Crucially, neither of these accounts are consistent with the increase in suppression we observe, which can only be explained by an increased recruitment of one overlapping population of neurons.



**Figure 7.1 Schematic representation of the phenomenon underlying an increase in repetition suppression with learning.** **A** Increase in suppression due to an increase in overlap between separate representations computing value for self and other in mPFC. **B** Increase in suppression due to a change in valuation for self in an agent-independent encoding of subjective value.

It is conceivable that value computations for self and other are performed within distinct populations of neurons, whose overlap increases due to prediction-error induced plasticity (Figure 7.1A). However, the effect could perhaps be explained even more parsimoniously if one assumes an agent-independent encoding of subjective value in mPFC (Nicolle et al., 2012), whereby the neural mechanisms involved in computing value for self and other are shared (Figure 7.1B). Subjective value could for example emerge from population codes computing a weighted sum over a distributed set of discounting units, each favouring a

certain discount rate (Kurth-Nelson and Redish, 2009). Such distributed encoding for different reward values in intertemporal choice has been observed in rat OFC (Roesch et al., 2006) and human mPFC (Wang et al., 2014). Learning-induced prediction errors could then act on the weights of the discounting units and an increase in overlap of the neural value representations for self and other would be directly related to a more similar subjective value computation (Figure 7.1B).

It has long been a matter of debate whether our remarkable social abilities are due to specially evolved brain regions, uniquely involved in social information processing (Amodio and Frith, 2006; Saxe, 2006), or whether the mechanisms underlying the computation of our own behaviour underlies our ability to infer other people's internal mental states. The notion of an agent-independent encoding of subjective value in mPFC is attractive, because it speaks to the idea that the same mechanisms that underlie the computation of our own goals and values could be used to model the goals and preferences of others (Buckner and Carroll, 2007; Mitchell, 2009).

### **7.3 Representing the structure of the world**

In Chapter 5, I show that the implicit exposure to an environment whose statistical transitions were determined by an underlying relational structure results in the formation of an abstract map in the hippocampal-entorhinal system. This suggests that the brain makes inferences about hidden structure in sensory data, and organizes this information in a map reminiscent of those maps supporting navigation in physical space. These studies address a fundamental question in neuroscience, namely whether the hippocampal-entorhinal system primarily computes spatial information or whether the same computations act on non-spatial knowledge (Buzsáki and Moser, 2013; Eichenbaum and Cohen, 2014; Tavares et al., 2015). Our findings suggests that a map is created within the hippocampal formation even in situations where relationships are non-spatial rather than spatial, discrete rather than continuous, and unavailable to conscious awareness. In physical space, the metric representation of relationships between landmarks allows for rapidly and flexibly computing distances and paths through space (Bush et al., 2015; Stemmler et al., 2015), enabling rapid rerouting when obstacles are introduced (Alvernhe et al., 2011) or removed (Alvernhe et al., 2008). Representing abstract relational knowledge using the same neural code greatly facilitates learning new problems in a complex world, and storing information in a map allows

for computing the relationships between items that have never been experienced together to facilitate decision making. However, it remains unclear how exactly such a relational map aids goal-directed behaviour.

One intriguing hypothesis proposes the hippocampal-entorhinal system stores an abstract cognitive map of the world, and orbitofrontal cortex (OFC) represents an animal's current location within this space (Wilson et al., 2014). OFC is critical for encoding stimulus-reward associations (Klein-Flügge et al., 2013b) and lesions in OFC lead to subtle impairments in specific decision making tasks such as reversal learning or devaluation paradigms. In reversal learning tasks animals initially learn to associate one of two stimuli or actions with a reward. After a number of trials the contingencies between a choice and a reward changes. OFC-lesioned animals have no difficulty in learning the initial task contingencies, but they are impaired in updating the contingencies after reversals (Butter, 1969). In devaluation paradigms, an animal's propensity to work for a reward is tested in a situation where the incentive value of the reward has been devalued. Unless animals are overtrained on the task they are typically sensitive to the devaluation procedure and stop working for the reward. OFC-lesioned animals do not display this sensitivity to reinforcer devaluation and continue working for a devalued reward (Gallagher et al., 1999). Successful performance on both types of tasks requires accurate credit assignment, whereby an outcome is specifically attributed to the relevant choice to guide behaviour (Walton et al., 2010). In other words, OFC encodes the currently relevant state of the world in order to guide adaptive decision making. A recent fMRI experiment in humans corroborates this hypothesis by demonstrating that an internal model of the association strength between stimuli and outcomes is stored in the hippocampus, but updated through OFC activity in situations where the association between stimuli and outcomes has to be updated flexibly (Boorman et al., 2016). This suggests that OFC does not encode a model of the entire task space per se, but instead represent only the behaviourally relevant state. OFC is particularly well suited for assigning a stimulus or an action to an outcome because its subdivisions are widely connected to brain areas processing multisensory and emotional information, memories and rewards (Kahnt et al., 2012). Direct access to a model representation stored in the hippocampus may be driven by the strong anatomical connections between the two anatomical structures (Carmichael and Price, 1995). However, understanding precisely how interactions between prefrontal cortex and hippocampus enable model-based behaviour is an important question for future research.



In the experiments reported in this thesis, I found no evidence for a representation of the relational structure in prefrontal cortex. It is conceivable that the reason for this is the incidental nature of the task. The entorhinal map I observe might be the consequence of latent learning, whereby animals construct a ‘cognitive map’ of the environment even in the absence of reinforcement (Tolman and Honzik, 1930). This map can then be used to enable rapid goal-directed learning when a reward is later introduced (Tolman, 1948). Future research should investigate whether elements of the relational map are represented in prefrontal cortex in situations where the map is useful for guiding goal-directed behaviour.

Another question these results raise is one addressing how exactly the brain knows which structure to represent in light of the large number of possible representations of the statistical relationships between elements in the environment. In Chapter 6, I demonstrate that in this specific situation the distance-dependence and the map-like representations of relational information can be reproduced by a simple Hopfield network with pairwise Hebbian plasticity between neighbouring stimulus representations, suggesting associative plasticity alone could lead to the formation of a map. As outlined in the discussion, however, the simultaneous representation of a memory and its relationship with other memories is unstable, ultimately resulting to a corrupted memory trace and a collapsed network.

A more efficient way of encoding information across experiences would be to reduce the dimensionality of the data by extracting the dimensions capturing most of the variance in the data. This is particularly relevant for the relationships between objects, events and other types of knowledge, which are not confined to a 2-dimensional space like the relationships between landmarks in physical space. In Chapter 6, I introduce a simple model of the relationship between place coding in the hippocampus and grid cells in entorhinal cortex, demonstrating that the typical hexagonal arrangement of grid cell firing could be explained by a covariance computation of hippocampal outputs. Critically, this relationship also results in a similar distance-dependent scaling of representational similarity, like the effect observed in the entorhinal cortex. The structure participants were trained on in the experimental Chapter 5 does not allow for differentiating between these two distinct accounts, and a higher-dimensional structure where the 2-dimensional topology is broken would need to be used to investigate this question. Importantly, such experiments might also provide new insights into computations performed by hippocampal place cells and entorhinal grid cells relevant for navigating physical space as it is currently still a matter of debate how place cells and grid cells encode 3-dimensional space. While whole-cell recordings of place cells reveal uniform

and nearly isotropic encoding of 3D space in bats flying freely in a volumetric space (Yartsev and Ulanovsky, 2013), place cells and grid cells in rats exploring 3-dimensional structures seem insensitive to the third dimension (Hayman et al., 2011).

## **7.4 Learning-induced acquisition and updating of world models**

A crucial aspect of learning about the world involves deciding when to modify an existing model, and when to form a new model of the environment (Gershman and Niv, 2010). The first set of studies reported in this thesis (Chapters 3-4) could be performed by modifying an existing model of an ‘unknown’ other person’s preferences, which is updated in light of new incoming information about the other’s choices. In the second set of experiments (Chapters 5-6), subjects constructed a completely novel representation of the relational structure of the environment. Under which conditions is it better to update an existing predictive model of the world, and when should a novel model be built?

While the question when to build a novel model has been underexplored, the issue of when to update an existing model of the world is comparably well understood. According to a ‘predictive coding’ theory of brain function, the statistics of the environment determine when to modify an existing representation of the world. ‘Predictive coding’ theorizes that the brain constantly compares incoming sensory information to signals that are predicted based upon our current model of the world. Events that are inconsistent with our world model elicit prediction errors (Friston, 2010), which are useful teaching signals because they can reflect a change in the environment and a need to update or replace our model in order to improve prediction of future events (O’Reilly et al., 2013). Changes in the representation of the brain’s model of the world are putatively driven by prediction error signals computed in VTA and substantia nigra, which directly influence memory-based models of the world by influencing hippocampal activity. Hippocampal and dopaminergic midbrain systems are intricately linked in a functional loop (Lisman and Grace, 2005) and the hippocampus itself receives significant dopaminergic innervation (Gasbarri et al., 1997) which influences long-term potentiation in CA1 (Morris et al., 2003). Information also flows in the other direction, whereby a hippocampal novelty signal is carried to the VTA resulting in a novelty-dependent dopamine release (Legault and Wise, 2001). More generally, the principle of ‘predictive

coding' thus also applies to models of memory function (Henson and Gagnepain, 2010) and human memory is well adapted to the statistics of the environment (Steyvers et al., 2006).

It has been suggested that a sequence of negative prediction errors, repeatedly elicited by the absence of predicted rewards as a consequence of altered statistics of the environment, can signal the necessity to create a novel world model (Redish et al., 2007). Gershman et al. (2010) extended this idea in a model of latent cause inference, whereby an animal constantly tries to predict the underlying causes of sensory data that are hidden from observation. Structure learning in this context is based on a generative model of the environment, and an animal needs to combine prior beliefs about how observations were generated with evidence provided by actual observations to infer the latent causes. Depending on the statistical structure of the environment and the size of the prediction errors, a model of the world is then either updated, or replaced (Gershman et al., 2014).

These observations are consistent with the phenomenon of remapping of cognitive maps (Muller and Kubie, 1987; Wilson and McNaughton, 1993). When an animal is exposed to an environment which gradually morphs from a square box to a circular box, place cells gradually change their firing fields (Leutgeb et al., 2005). This manipulation presumably introduces only small prediction errors, resulting in a gradual update of the model of the world. If, however, morphs are presented at random rather than in consecutive order, the cognitive map abruptly and coherently changes its representation from a circle-like to a squarelike attractor state (Wills et al., 2005), suggesting that the animal fundamentally updates its belief about the current state of the world. This phenomenon also addresses the fundamental distinction between pattern completion and pattern separation, which needs to trade off situations where a new memory is encoded, or an old memory is retrieved. The properties of the hippocampus minimize the trade-off posed by assigning an activity pattern to an existing memory or a new memory (O'Reilly and McClelland, 1994). How precisely the brain trades off the decision to update an existing model of the world, or to construct a new one is an important question for future research.

## 7.5 Conclusion

In this thesis, I present a series of studies investigating the neural mechanisms underlying learning, memory formation and choice. Using fMRI repetition suppression and computational modelling, I demonstrate (1) how prediction errors induced by learning about

the environment result in plasticity in a value computation network. This plasticity can account for a malleability of subjective preferences. Furthermore, I show (2) how associations between objects in the world are combined into a model of the world and stored in the hippocampal-entorhinal system as a cognitive map. Such maps can form the basis of goal-directed behaviour, because they allow for inferring relationships without direct experience. Both sets of studies demonstrate how information is represented and updated at the level of neural representations, providing a bridge between representational codes and cognitive computations.

---

## REFERENCES

- Abbott, L.F., Varela, J.A., Sen, K., and Nelson, S.B. (1997). Synaptic depression and cortical gain control. *Science* *275*, 220–224.
- Abe, K., and Watanabe, D. (2011). Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nat. Neurosci.* *14*, 1067–1074.
- Ainge, J.A., Tamosiunaite, M., Woergoetter, F., and Dudchenko, P.A. (2007). Hippocampal CA1 place cells encode intended destination on a maze with multiple choice points. *J. Neurosci.* *27*, 9769–9779.
- Albrecht, D.G., Farrar, S.B., and Hamilton, D.B. (1984). Spatial contrast adaptation characteristics of neurones recorded in the cat’s visual cortex. *J. Physiol.* *347*, 713–739.
- Alessi, S.M., and Petry, N.M. (2003). Pathological gambling severity is associated with impulsivity in a delay discounting procedure. *Behav. Processes* *64*, 345–354.
- Allen, T.A., Salz, D.M., McKenzie, S., and Fortin, N.J. (2016). Nonspatial sequence coding in CA1 neurons. *J. Neurosci.* *36*, 1547–1563.
- Alvernhe, A., Van Cauter, T., Save, E., and Poucet, B. (2008). Different CA1 and CA3 representations of novel routes in a shortcut situation. *J. Neurosci.* *28*, 7324–7333.
- Alvernhe, A., Save, E., and Poucet, B. (2011). Local remapping of place cell firing in the Tolman detour task. *Eur. J. Neurosci.* *33*, 1696–1705.
- Amarimber, S.-I., and Amari, S. (1972). Characteristics of random nets of analog neuron-like elements. *IEEE Trans. Syst. Man. Cybern.* *2*, 643–657.
- Amodio, D., and Frith, C. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nat. Rev. Neurosci.*
- Andersson, J.L., Hutton, C., Ashburner, J., Turner, R., and Friston, K. (2001). Modeling geometric deformations in EPI time series. *Neuroimage* *13*, 903–919.
- Arthurs, O.J., and Boniface, S. (2002). How well do we understand the neural origins of the fMRI BOLD signal? *Trends Neurosci.* *25*, 27–31.
- Ashby, F.G., and Maddox, W.T. (2005). Human category learning. *Annu. Rev. Psychol.* *56*, 149–178.
- Atlas, L.Y., Bolger, N., Lindquist, M.A., and Wager, T.D. (2010). Brain mediators of predictive cue effects on perceived pain. *J. Neurosci.* *30*, 12964–12977.
- Auksztulewicz, R., and Friston, K. (2016). Repetition suppression and its contextual determinants in predictive coding. *Cortex*.
- Averbeck, B.B., Latham, P.E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* *7*, 358–366.
- Baene, W. De, and Vogels, R. (2010). Effects of Adaptation on the Stimulus Selectivity of Macaque Inferior Temporal Spiking Activity and Local Field Potentials. *Cereb. Cortex* *20*,

---

2145–2165.

Barrash, J., Tranel, D., and Anderson, S.W. (2000). Acquired personality disturbances associated with bilateral damage to the ventromedial prefrontal region. *Dev. Neuropsychol.* *18*, 355–381.

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

Barron, H.C., Vogels, T.P., Emir, U.E., Makin, T.R., O’Shea, J., Clare, S., Jbabdi, S., Dolan, R.J., and Behrens, T.E.J. (2016a). Unmasking latent inhibitory connections in human cortex to reveal dormant cortical memories. *Neuron* *90*, 191–203.

Barron, H.C., Garvert, M.M., and Behrens, T.E.J. (2016b). Repetition suppression: a means to index neural representations using BOLD? *Philos. Trans. R. Soc. B Biol. Sci.* *371*, 20150355.

Barry, C., Lever, C., Hayman, R., Hartley, T., Burton, S., O’Keefe, J., Jeffery, K., and Burgess, N. (2006). The boundary vector cell model of place cell firing and spatial memory. *Rev. Neurosci.* *17*, 71–97.

Barry, C., Hayman, R., Burgess, N., and Jeffery, K.J. (2007). Experience-dependent rescaling of entorhinal grids. *Nat. Neurosci.* *10*, 682–684.

Bartfeld, E., and Grinvald, A. (1992). Relationships between orientation-preference pinwheels, cytochrome oxidase blobs, and ocular-dominance columns in primate striate cortex. *Proc. Natl. Acad. Sci. U. S. A.* *89*, 11905–11909.

Bartra, O., McGuire, J.T., and 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.

Bastos, A.M., Vezoli, J., Bosman, C.A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J.R., De Weerd, P., Kennedy, H., and Fries, P. (2015). Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels. *Neuron* *85*, 390–401.

Baylis, G.C., and Rolls, E.T. (1987). Responses of neurons in the inferior temporal cortex in short term and serial recognition memory tasks. *Exp. Brain Res.* *65*, 614–622.

Bechara, A. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain* *123*, 2189–2202.

Bechara, A., Damasio, A.R., Damasio, H., and Anderson, S.W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* *50*, 7–15.

Behrens, T.E.J., Woolrich, M.W., Walton, M.E., and Rushworth, M.F.S. (2007). Learning the value of information in an uncertain world. *Nat. Neurosci.* *10*, 1214–1221.

Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., and Rushworth, M.F.S. (2008). Associative learning of social value. *Nature* *456*, 245–249.

Behrens, T.E.J., Hunt, L.T., and Rushworth, M.F.S. (2009). The computation of social behavior. *Science* *324*, 1160–1164.

Berns, G.S., Capra, C.M., Moore, S., and Noussair, C. (2010). Neural mechanisms of the influence of popularity on adolescent ratings of music. *Neuroimage* *49*, 2687–2696.

Binder, J., Desai, R., Graves, W., and Conant, L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex.*

Birn, R.M., Diamond, J.B., Smith, M.A., and Bandettini, P.A. (2006). Separating respiratory-

- 
- variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. *Neuroimage* *31*, 1536–1548.
- Blakemore, C., and Campbell, F.W. (1969). Adaptation to spatial stimuli. *J. Physiol.* *200*, 11P–13P.
- Bliss, T.V.P., and Lømo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *J. Physiol.* *232*, 331–356.
- Bloch, F. (1946). Nuclear Induction. *Phys. Rev.* *70*, 460–474.
- Blum, K.I., and Abbott, L.F. (1996). A Model of Spatial Map Formation in the Hippocampus of the Rat. *Neural Comput.* *8*, 85–93.
- Bond, A.B., Kamil, A.C., and Balda, R.P. (2003). Social complexity and transitive inference in corvids. *Anim. Behav.* *65*, 479–487.
- Bonnevie, T., Dunn, B., Fyhn, M., Hafting, T., Derdikman, D., Kubie, J.L., Roudi, Y., Moser, E.I., and 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., and 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., Rajendrana, V.G., O'Reilly, J.X., and Behrens, T.E.J. (2016). Two anatomically and computationally distinct learning signals predict changes to stimulus-outcome associations in hippocampus. *Neuron* *89*, 1343–1354.
- Bouret, S., and Richmond, B.J. (2010). Ventromedial and orbital prefrontal neurons differentially encode internally and externally driven motivational values in monkeys. *J. Neurosci.* *30*, 8591–8601.
- Boyd, R., Richerson, P.J., and Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proc. Natl. Acad. Sci.* *108*, 10918–10925.
- Brun, V.H., Otnass, M.K., Molden, S., Steffenach, H.-A., Witter, M.P., Moser, M.-B., and Moser, E.I. (2002). Place cells and place recognition maintained by direct entorhinal-hippocampal circuitry. *Science* *296*, 2243–2246.
- Buckner, R.L., and Carroll, D.C. (2007). Self-projection and the brain. *Trends Cogn. Sci.* *11*, 49–57.
- Buckner, R.L., Goodman, J., Burock, M., Rotte, M., Koutstaal, W., Schacter, D., Rosen, B., and Dale, A.M. (1998). Functional-anatomic correlates of object priming in humans revealed by rapid presentation event-related fMRI. *Neuron* *20*, 285–296.
- Bunsey, M., and Eichenbaum, H. (1996). Conservation of hippocampal memory function in rats and humans. *Nature* *379*, 255–257.
- Bush, D., Barry, C., and Burgess, N. (2014). What do grid cells contribute to place cell firing? *Trends Neurosci.* *37*, 136–145.
- Bush, D., Barry, C., Manson, D., and Burgess, N. (2015). Using grid cells for navigation. *Neuron* *87*, 507–520.
- Butter, C. (1969). Perseveration in extinction and in discrimination reversal tasks following selective frontal ablations in *Macaca mulatta*. *Physiol. Behav.*

- 
- Buzsáki, G., and Moser, E.I. (2013). Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nat. Neurosci.* *16*, 130–138.
- Calabresi, P., Picconi, B., Tozzi, A., and Di Filippo, M. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends Neurosci.* *30*, 211–219.
- Campbell-Meiklejohn, D.K., Bach, D.R., Roepstorff, A., Dolan, R.J., and Frith, C.D. (2010). How the opinion of others affects our valuation of objects. *Curr. Biol.* *20*, 1165–1170.
- Carandini, M., and Ferster, D. (1997). A Tonic Hyperpolarization Underlying Contrast Adaptation in Cat Visual Cortex. *Science* (80-. ). *276*, 949–952.
- Carmichael, S.T., and Price, J.L. (1995). Limbic connections of the orbital and medial prefrontal cortex in macaque monkeys. *J. Comp. Neurol.* *363*, 615–641.
- Chadwick, M.J., Jolly, A.E.J., Amos, D.P., Hassabis, D., and Spiers, H.J. (2015). A goal direction signal in the human entorhinal/subicular region. *Curr. Biol.* *25*, 87–92.
- Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., and Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature* *487*, 51–56.
- Cialdini, R.B.R., and Goldstein, N.J.N. (2004). Social influence: Compliance and conformity. *Annu. Rev. Psychol.* *55*, 591–621.
- Clithero, J.A., and Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. *Soc. Cogn. Affect. Neurosci.* *9*, 1289–1302.
- Cohen-Kashi Malina, K., Jubran, M., Katz, Y., and Lampl, I. (2013). Imbalance between excitation and inhibition in the somatosensory cortex produces postadaptation facilitation. *J. Neurosci.* *33*, 8463–8471.
- Collin, S.H.P., Milivojevic, B., and Doeller, C.F. (2015). Memory hierarchies map onto the hippocampal long axis in humans. *Nat. Neurosci.* *18*, 1562–1564.
- Condé, F., Maire-Lepoivre, E., Audinat, E., and Crépel, F. (1995). Afferent connections of the medial frontal cortex of the rat. II. Cortical and subcortical afferents. *J. Comp. Neurol.* *352*, 567–593.
- Constantinescu, A.O., O’Reilly, J.X., and Behrens, T.E.J. Organizing conceptual knowledge in humans with a grid-like code.
- Cooper, N., Kable, J.W., Kim, B.K., and Zauberman, G. (2013). Brain activity in valuation regions while thinking about the future predicts individual discount rates. *J. Neurosci.* *33*, 13150–13156.
- Cowansage, K.K., Shuman, T., Dillingham, B.C., Chang, A., Golshani, P., and Mayford, M. (2014). Direct reactivation of a coherent neocortical memory of context. *Neuron* *84*, 432–441.
- Csicsvari, J., Hirase, H., Mamiya, A., and Buzsáki, G. (2000). Ensemble patterns of hippocampal CA3-CA1 neurons during sharp wave-associated population events. *Neuron* *28*, 585–594.
- Czurkó, A., Hirase, H., Csicsvari, J., and Buzsáki, G. (1999). Sustained activation of hippocampal pyramidal cells by “space clamping” in a running wheel. *Eur. J. Neurosci.* *11*, 344–352.
- D’Ardenne, K., McClure, S.M., Nystrom, L.E., and Cohen, J.D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science* *319*, 1264–1267.
- D’Argembeau, A., Ruby, P., Collette, F., Degueldre, C., Balteau, E., Luxen, A., Maquet, P.,



- 
- and Salmon, E. (2007). Distinct regions of the medial prefrontal cortex are associated with self-referential processing and perspective taking. *J. Cogn. Neurosci.* *19*, 935–944.
- Davidson, T.J., Kloosterman, F., and Wilson, M.A. (2009). Hippocampal replay of extended experience. *Neuron* *63*, 497–507.
- Daw, N.D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* *8*, 1704–1711.
- Daw, N.D., O’Doherty, J.P., Dayan, P., Seymour, B., and Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. *Nature* *441*, 876–879.
- Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., and Dolan, R.J. (2011). Model-Based Influences on Humans’ Choices and Striatal Prediction Errors. *Neuron* *69*, 1204–1215.
- Dayan, P. (1993). Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Comput.* *5*, 613–624.
- Dayan, P., and Balleine, B.W. (2002). Reward, Motivation, and Reinforcement Learning. *Neuron* *36*, 285–298.
- Dayan, P., and Berridge, K. (2014). Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cogn. Affect. Behav. Neurosci.*
- Denny, B.T., Kober, H., Wager, T.D., and Ochsner, K.N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *J. Cogn. Neurosci.* *24*, 1742–1752.
- Derdikman, D., and Moser, E.I. (2010). A manifold of spatial maps in the brain. *Trends Cogn. Sci.* *14*, 561–569.
- Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proc. Natl. Acad. Sci. U. S. A.* *93*, 13494–13499.
- Deutsch, M., and Gerard, H. (1955). A study of normative and informational social influences upon individual judgment. *J. Abnorm. Soc.* ....
- Diba, K., and Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nat. Neurosci.* *10*, 1241–1242.
- Doeller, C.F., Barry, C., and Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature* *463*, 657–661.
- Doll, B.B., Duncan, K.K.D., Simon, D.D.A., Shohamy, D., and Daw, N.D. (2015). Model-based choices involve prospective neural activity. *Nat. Neurosci.* *18*, 767–772.
- Dombeck, D.A., Harvey, C.D., Tian, L., Looger, L.L., and Tank, D.W. (2010). Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. *Nat. Neurosci.* *13*, 1433–1440.
- Dordek, Y., Soudry, D., Meir, R., and Derdikman, D. (2016). Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *Elife* *5*, e10094.
- Downing, P.E., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science* *293*, 2470–2473.
- Drucker, D.M., and Aguirre, G.K. (2009). Different Spatial Scales of Shape Similarity Representation in Lateral and Ventral LOC. *Cereb. Cortex* *19*, 2269–2280.

- 
- Dunbar, R.I.M., and Shultz, S. (2007). Evolution in the social brain. *Science* 317, 1344–1347.
- Dupret, D., O’Neill, J., Pleydell-Bouverie, B., and Csicsvari, J. (2010). The reorganization and reactivation of hippocampal maps predict spatial memory performance. *Nat. Neurosci.* 13, 995–1002.
- Edelson, M., Sharot, T., Dolan, R.J., and Dudai, Y. (2011). Following the crowd: Brain substrates of long-term memory conformity. *Science* 333, 108–111.
- Eden, C.G., Lamme, V.A.F., and Uylings, H.B.M. (1992). Heterotopic cortical afferents to the medial prefrontal cortex in the rat. A combined retrograde and anterograde tracer study. *Eur. J. Neurosci.* 4, 77–97.
- Egner, T., Monti, J.M., and Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *J. Neurosci.* 30, 16601–16608.
- Eichenbaum, H., and Cohen, N.J. (2014). Can we reconcile the declarative memory and spatial navigation views on hippocampal function? *Neuron* 83, 764–770.
- Ekstrom, A.D., Kahana, M.J., Caplan, J.B., Fields, T.A., Isham, E.A., Newman, E.L., and Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature* 425, 184–188.
- Epstein, R., and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature* 392, 598–601.
- Epstein, R.A., Parker, W.E., and Feiler, A.M. (2007). Where am I now? Distinct roles for parahippocampal and retrosplenial cortices in place recognition. *J. Neurosci.* 27, 6141–6149.
- Ersner-Hershfield, H., Wimmer, G.E., and Knutson, B. (2009). Saving for the future self: Neural measures of future self-continuity predict temporal discounting. *Soc. Cogn. Affect. Neurosci.* 4, 85–92.
- Euston, D.R., Gruber, A.J., and McNaughton, B.L. (2012). The Role of Medial Prefrontal Cortex in Memory and Decision Making. *Neuron* 76, 1057–1070.
- Evenden, J.L. (1999). Varieties of impulsivity. *Psychopharmacology (Berl.)* 146, 348–361.
- Ezzyat, Y., and Davachi, L. (2014). Similarity breeds proximity: pattern similarity within and across contexts is related to later mnemonic judgments of temporal proximity. *Neuron* 81, 1179–1189.
- Falk, E.B., Berkman, E.T., Mann, T., Harrison, B., and Lieberman, M.D. (2010). Predicting persuasion-induced behavior change from the brain. *J. Neurosci.* 30, 8421–8424.
- Fechner, G.T. (1860). Elements of psychophysics, 1860. In *Readings in the History of Psychology*, W. Dennis, ed.
- Fell, J., Klaver, P., Lehnertz, K., Grunwald, T., Schaller, C., Elger, C.E., and Fernández, G. (2001). Human memory formation is accompanied by rhinal–hippocampal coupling and decoupling. *Nat. Neurosci.* 4, 1259–1264.
- Ferbinteanu, J., and Shapiro, M.L. (2003). Prospective and Retrospective Memory Coding in the Hippocampus. *Neuron* 40, 1227–1239.
- Ferrand, L., and Grainger, J. (1992). Phonology and orthography in visual word recognition: evidence from masked non-word priming. *Q. J. Exp. Psychol. A.* 45, 353–372.
- Ferster, C., and Skinner, B. (1957). Schedules of reinforcement.
- Fiorillo, C.D., Tobler, P.N., and Schultz, W. (2003). Discrete coding of reward probability

- 
- and uncertainty by dopamine neurons. *Science* 299, 1898–1902.
- FitzGerald, T.H.B., Seymour, B., and Dolan, R.J. (2009). The Role of Human Orbitofrontal Cortex in Value Comparison for Incommensurable Objects. *J. Neurosci.* 29, 8388–8395.
- Foster, D.J., and Wilson, M.A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* 440, 680–683.
- Foster, T., Castro, C., and McNaughton, B. (1989). Spatial selectivity of rat hippocampal neurons: dependence on preparedness for movement. *Science* 244, 1580–1582.
- Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. London B Biol. Sci.* 360, 815–836.
- Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301.
- Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138.
- Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., and Frackowiak, R.S.J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* 2, 189–210.
- Friston, K.J., Frith, C.D., Turner, R., and Frackowiak, R.S. (1995). Characterizing evoked hemodynamics with fMRI. *Neuroimage* 2, 157–165.
- Friston, K.J., Zarahn, E., Josephs, O., Henson, R.N.A., and Dale, A.M. (1999). Stochastic designs in event-related fMRI. *Neuroimage* 10, 607–619.
- Frith, U., and Frith, C. (2010). The social brain: allowing humans to boldly go where no other species has been. *Philos. Trans. R. Soc. B Biol. Sci.* 365, 165–176.
- Fuhs, M., and Touretzky, D. (2006). A spin glass model of path integration in rat medial entorhinal cortex. *J. Neurosci.*
- Fyhn, M., Hafting, T., Treves, A., Moser, M.-B., and Moser, E.I. (2007). Hippocampal remapping and grid realignment in entorhinal cortex. *Nature* 446, 190–194.
- Gallagher, H.L., and Frith, C.D. (2003). Functional imaging of “theory of mind.” *Trends Cogn. Sci.* 7, 77–83.
- Gallagher, M., McMahan, R.W., and Schoenbaum, G. (1999). Orbitofrontal Cortex and Representation of Incentive Value in Associative Learning. *J. Neurosci.* 19, 6610–6614.
- Garrido, M.I., Kilner, J.M., Kiebel, S.J., and Friston, K.J. (2007). Evoked brain responses are generated by feedback loops. *Proc. Natl. Acad. Sci. U. S. A.* 104, 20961–20966.
- Garris, P.A., Collins, L.B., Jones, S.R., and Wightman, R.M. (2006). Evoked Extracellular Dopamine In Vivo in the Medial Prefrontal Cortex. *J. Neurochem.* 61, 637–647.
- Gasbarri, A., Sulli, A., and Packard, M.G. (1997). The dopaminergic mesencephalic projections to the hippocampal formation in the rat. *Prog. Neuro-Psychopharmacology Biol. Psychiatry* 21, 1–22.
- Gentry, G., Brown, W.L., Kaplan, S.J., and Kaplan, S.J. (1947). An experimental analysis of the spatial location hypothesis in learning. *J. Comp. Physiol. Psychol.* 40, 309–322.
- Gershman, S.J., and Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Curr. Opin. Neurobiol.* 20, 251–256.
- Gershman, S.J., Blei, D.M., and Niv, Y. (2010). Context, learning, and extinction. *Psychol.*

---

Rev. 117, 197–209.

Gershman, S.J., Radulescu, A., Norman, K.A., and Niv, Y. (2014). Statistical computations underlying the dynamics of memory updating. *PLoS Comput. Biol.* 10, e1003939.

Girardeau, G., Benchenane, K., Wiener, S.I., Buzsáki, G., and Zugaro, M.B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. *Nat. Neurosci.* 12, 1222–1223.

Gläscher, J., Daw, N., Dayan, P., and O’Doherty, J.P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron* 66, 585–595.

Glover, G.H., Li, T.Q., and Ress, D. (2000). Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magn. Reson. Med. Off. J. Soc. Magn. Reson. Med. / Soc. Magn. Reson. Med.* 44, 162–167.

Gomperts, S.N., Kloosterman, F., and Wilson, M.A. (2015). VTA neurons coordinate with the hippocampal reactivation of spatial experience. *Elife* 4, e05360.

Green, J., and Arduini, A. (1954). Hippocampal electrical activity in arousal. *J. Neurophysiol.*

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzhak, Y., and Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron* 24, 187–203.

Grill-Spector, K., Henson, R., and Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn. Sci.* 10, 14–23.

Griniasty, M., Tsodyks, M. V., and Amit, D.J. (1993). Conversion of Temporal Correlations Between Stimuli to Spatial Correlations Between Attractors. *Neural Comput.* 5, 1–17.

Grosenick, L., Clement, T.S., and Fernald, R.D. (2007). Fish can infer social rank by observation alone. *Nature* 445, 429–432.

Gross, C.G., Schiller, P.H., Wells, C., and Gerstein, G.L. (1967). Single-unit activity in temporal association cortex of the monkey. *J. Neurophysiol.* 30, 833–843.

Gross, C.G., Bender, D.B., and Rocha-Miranda, C.E. (1969). Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science* 166, 1303–1306.

Haber, S.N., and 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., and Moser, E.I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature* 436, 801–806.

Hafting, T., Fyhn, M., Bonnevie, T., Moser, M.-B., and Moser, E.I. (2008). Hippocampus-independent phase precession in entorhinal grid cells. *Nature* 453, 1248–1252.

Hales, J.B., Schlesiger, M.I., Leutgeb, J.K., Squire, L.R., Leutgeb, S., and Clark, R.E. (2014). Medial entorhinal cortex lesions only partially disrupt hippocampal place cells and hippocampus-dependent place memory. *Cell Rep.* 9, 893–901.

Hampton, A.N., Bossaerts, P., and 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., Malmaud, J., and Rangel, A. (2011). Focusing attention on the health aspects of foods changes value signals in vmPFC and improves dietary choice. *J. Neurosci.* 31, 11077–11087.

- 
- Harlow, J. (1868). Recovery from the passage of an iron bar through the head. *Publ. Massachusetts Med. Soc.*
- Harris, L.T., and Fiske, S.T. (2010). Neural regions that underlie reinforcement learning are also active for social expectancy violations. *Soc. Neurosci.* *5*, 76–91.
- Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *Neuroimage* *62*, 852–855.
- Haxby, J. V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., and Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science* (80-.). *293*, 2425–2430.
- Hayden, B.Y., Heilbronner, S.R., Pearson, J.M., and Platt, M.L. (2011). Surprise signals in anterior cingulate cortex: Neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *J. Neurosci.* *31*, 4178–4187.
- Hayman, R., Verriotis, M.A., Jovalekic, A., Fenton, A.A., and Jeffery, K.J. (2011). Anisotropic encoding of three-dimensional space by place cells and grid cells. *Nat. Neurosci.* *14*, 1182–1188.
- Haynes, J.-D., and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* *7*, 523–534.
- Hebb, O. (1949). *The organization of behavior: A neuropsychological approach* (John Wiley & Sons).
- Heckers, S., Zalesak, M., Weiss, A.P., Ditman, T., and Titone, D. (2004). Hippocampal activation during transitive inference in humans. *Hippocampus* *14*, 153–162.
- Henson, R.N., and Gagnepain, P. (2010). Predictive, interactive multiple memory systems. *Hippocampus* *20*, 1315–1326.
- Henson, R., Shallice, T., and Dolan, R. (2000). Neuroimaging evidence for dissociable forms of repetition priming. *Science* (80-.). *287*, 1269–1272.
- Henze, D.A., Wittner, L., and Buzsáki, G. (2002). Single granule cells reliably discharge targets in the hippocampal CA3 network in vivo. *Nat. Neurosci.* *5*, 790–795.
- Hershberger, W.A. (1986). An approach through the looking-glass. *Anim. Learn. Behav.* *14*, 443–451.
- Hiramatsu, C., Goda, N., and Komatsu, H. (2011). Transformation from image-based to perceptual representation of materials along the human ventral visual pathway. *Neuroimage* *57*, 482–494.
- Hoffman, K.L., and McNaughton, B.L. (2002). Coordinated Reactivation of Distributed Memory Traces in Primate Neocortex. *Science* (80-.). *297*, 2070–2073.
- Hok, V., Lenck-Santini, P.-P., Roux, S., Save, E., Muller, R.U., and Poucet, B. (2007). Goal-related activity in hippocampal place cells. *J. Neurosci.* *27*, 472–482.
- Holland, P.C. *Relations Between Pavlovian-Instrumental Transfer and Reinforcer Devaluation.*
- Hollup, S.A., Molden, S., Donnett, J.G., Moser, M.-B., and Moser, E.I. (2001). Accumulation of Hippocampal Place Fields at the Goal Location in an Annular Watermaze Task. *J. Neurosci.* *21*, 1635–1644.
- Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci.* *79*, 2554–2558.

- 
- Horner, A.J., Bisby, J.A., Bush, D., Lin, W.-J., and Burgess, N. (2015). Evidence for holistic episodic recollection via hippocampal pattern completion. *Nat. Commun.* *6*, 7462.
- Howard, J.D., Gottfried, J.A., Tobler, P.N., and Kahnt, T. (2015). Identity-specific coding of future rewards in the human orbitofrontal cortex. *Proc. Natl. Acad. Sci. U. S. A.* *112*, 5195–5200.
- Howard, L.R., Javadi, A.H., Yu, Y., Mill, R.D., Morrison, L.C., Knight, R., Loftus, M.M., Staskute, L., and Spiers, H.J. (2014). The Hippocampus and Entorhinal Cortex Encode the Path and Euclidean Distances to Goals during Navigation. *Curr. Biol.* *24*, 1331–1340.
- Hsieh, L.-T., Gruber, M.J., Jenkins, L.J., and Ranganath, C. (2014). Hippocampal activity patterns carry information about objects in temporal context. *Neuron* *81*, 1165–1178.
- Hunt, L.T., Kolling, N., Soltani, A., Woolrich, M.W., Rushworth, M.F.S., and Behrens, T.E.J. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nat. Neurosci.* *15*, 470–476.
- Hutton, C., Josephs, O., Stadler, J., Featherstone, E., Reid, A., Speck, O., Bernarding, J., and Weiskopf, N. (2011). The impact of physiological noise correction on fMRI at 7 T. *Neuroimage* *57*, 101–112.
- Hyman, J.M., Ma, L., Balaguer-Ballester, E., Durstewitz, D., and Seamans, J.K. (2012). Contextual encoding by ensembles of medial prefrontal cortex neurons. *Proc. Natl. Acad. Sci.* *109*, 5086–5091.
- Insausti, R. (1993). Comparative anatomy of the entorhinal cortex and hippocampus in mammals. *Hippocampus* *3*, 19–26.
- Jackson, P.L., Meltzoff, A.N., and Decety, J. (2005). How do we perceive the pain of others? A window into the neural processes involved in empathy. *Neuroimage* *24*, 771–779.
- Jacobs, B. (2001). Regional Dendritic and Spine Variation in Human Cerebral Cortex: a Quantitative Golgi Study. *Cereb. Cortex* *11*, 558–571.
- Jacobs, J., Weidemann, C.T., Miller, J.F., Solway, A., Burke, J.F., Wei, X.-X., Suthana, N., Sperling, M.R., Sharan, A.D., Fried, I., et al. (2013). Direct recordings of grid-like neuronal activity in human spatial navigation. *Nat. Neurosci.* *16*, 1188–1190.
- Jay, T.M., and Witter, M.P. (1991). Distribution of hippocampal CA1 and subicular efferents in the prefrontal cortex of the rat studied by means of anterograde transport of Phaseolus vulgaris-leucoagglutinin. *J. Comp. Neurol.* *313*, 574–586.
- Jenkins, A.C., Macrae, C.N., and Mitchell, J.P. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proc. Natl. Acad. Sci.* *105*, 4507–4512.
- Jetten, J., Hornsey, M.J., and Adarves-Yorno, I. (2006). When group members admit to being conformist: the role of relative intragroup status in conformity self-reports. *Pers. Soc. Psychol. Bull.* *32*, 162–173.
- Ji, D., and Wilson, M.A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat. Neurosci.* *10*, 100–107.
- Johnson, A., and 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, M.W., and Wilson, M.A. (2005). Phase precession of medial prefrontal cortical activity relative to the hippocampal theta rhythm. *Hippocampus* *15*, 867–873.

- 
- Kable, J.W., and Glimcher, P.W. (2007). The neural correlates of subjective value during intertemporal choice. *Nat Neurosci* *10*, 1625–1633.
- Kahnt, T., Heinzle, J., Park, S.Q., and Haynes, J.-D. (2010). The neural code of reward anticipation in human orbitofrontal cortex. *Proc. Natl. Acad. Sci. U. S. A.* *107*, 6010–6015.
- Kahnt, T., Chang, L.J., Park, S.Q., Heinzle, J., and Haynes, J.-D. (2012). Connectivity-based parcellation of the human orbitofrontal cortex. *J. Neurosci.* *32*, 6240–6250.
- Kamin, L. (1969). Predictability, surprise, attention, and conditioning. *Punishm. Aversive Behav.*
- Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* *17*, 4302–4311.
- Karlsson, M.P., and Frank, L.M. (2009). Awake replay of remote experiences in the hippocampus. *Nat. Neurosci.* *12*, 913–918.
- Kelley, W.M., Macrae, C.N., Wyland, C.L., Caglar, S., Inati, S., and Heatherton, T.F. (2002). Finding the self? An event-related fMRI study. *J. Cogn. Neurosci.* *14*, 785–794.
- Kilner, J.M., Neal, A., Weiskopf, N., Friston, K.J., and Frith, C.D. (2009). Evidence of mirror neurons in human inferior frontal gyrus. *J. Neurosci.* *29*, 10153–10159.
- King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R., and Montague, P.R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science* *308*, 78–83.
- Kirby, K.N. (2009). One-year temporal stability of delay-discount rates. *Psychon. Bull. Rev.* *16*, 457–462.
- Kjelstrup, K.B., Solstad, T., Brun, V.H., Hafting, T., Leutgeb, S., Witter, M.P., Moser, E.I., and Moser, M.-B. (2008). Finite scale of spatial representation in the hippocampus. *Science* *321*, 140–143.
- Klein-Flügge, M.C., Hunt, L.T., Bach, D.R., Dolan, R.J., and Behrens, T.E.J. (2011). Dissociable reward and timing signals in human midbrain and ventral striatum. *Neuron* *72*, 654–664.
- Klein-Flügge, M.C., Barron, H.C., Brodersen, K.H., Dolan, R.J., and Behrens, T.E.J. (2013a). Segregated Encoding of Reward–Identity and Stimulus–Reward Associations in Human Orbitofrontal Cortex. *J. Neurosci.* *33*, 3202–3211.
- Klein-Flügge, M.C.C., Barron, H.C.C., Brodersen, K.H.H., Dolan, R.J.J., and Behrens, T.E.J.E.J. (2013b). Segregated encoding of reward-identity and stimulus-reward associations in human orbitofrontal cortex. *J. Neurosci.* *33*, 3202–3211.
- Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., and Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron* *61*, 140–151.
- Koenig, J., Linder, A.N., Leutgeb, J.K., and Leutgeb, S. (2011). The spatial periodicity of grid cells is not sustained during reduced theta oscillations. *Science* *332*, 592–595.
- Kohn, A. (2007). Visual adaptation: Physiology, mechanisms, and functional benefits. *J. Neurophysiol.* *97*, 3155–3164.
- Komorowski, R.W., Garcia, C.G., Wilson, A., Hattori, S., Howard, M.W., and Eichenbaum, H. (2013). Ventral hippocampal neurons are shaped by experience to represent behaviorally relevant contexts. *J. Neurosci.* *33*, 8079–8087.
- Kourtzi, Z., and Kanwisher, N. (2001). Representation of perceived object shape by the

- 
- human lateral occipital complex. *Science* (80-.). *293*, 1506–1509.
- Kreek, M.J., Nielsen, D.A., Butelman, E.R., and LaForge, K.S. (2005). Genetic influences on impulsivity, risk taking, stress responsivity and vulnerability to drug abuse and addiction. *Nat. Neurosci.* *8*, 1450–1457.
- Krekelberg, B., Boynton, G.M., and van Wezel, R.J.A. (2006). Adaptation: from single cells to BOLD signals. *Trends Neurosci.* *29*, 250–256.
- Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* *17*, 401–412.
- Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* *2*, 4.
- Krienen, F.M., Tu, P.-C., and Buckner, R.L. (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. *J. Neurosci.* *30*, 13906–13915.
- Krupic, J., Bauza, M., Burton, S., Lever, C., and O’Keefe, J. (2014). How environment geometry affects grid cell symmetry and what we can learn from it. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* *369*, 20130188.
- Kumaran, D., Melo, H.L., and Duzel, E. (2012). The Emergence and Representation of Knowledge about Social and Nonsocial Hierarchies. *Neuron* *76*, 653–666.
- Kurth-Nelson, Z., and Redish, A.D. (2009). Temporal-difference reinforcement learning with distributed representations. *PLoS One* *4*, e7362.
- Langston, R.F., Ainge, J.A., Couey, J.J., Canto, C.B., Bjerknes, T.L., Witter, M.P., Moser, E.I., and Moser, M.-B. (2010). Development of the spatial representation system in the rat. *Science* *328*, 1576–1580.
- Larsson, J., Solomon, S.G., and Kohn, A. (2015). fMRI adaptation revisited. *Cortex*.
- Lauritzen, M., and Gold, L. (2003). Brain Function and Neurophysiological Correlates of Signals Used in Functional Neuroimaging. *J. Neurosci.* *23*, 3972–3980.
- Lavenex, P., and Amaral, D. (2000). Hippocampal-neocortical interaction: a hierarchy of associativity. *Hippocampus*.
- de Lavilléon, G., Lacroix, M.M., Rondi-Reig, L., and Benchenane, K. (2015). Explicit memory creation during sleep demonstrates a causal role of place cells in navigation. *Nat. Neurosci.* *18*, 493–495.
- Lebreton, M., Jorge, S., Michel, V., Thirion, B., and Pessiglione, M. (2009). An automatic valuation system in the human brain: Evidence from functional neuroimaging. *Neuron* *64*, 431–439.
- Legault, M., and Wise, R.A. (2001). Novelty-evoked elevations of nucleus accumbens dopamine: dependence on impulse flow from the ventral subiculum and glutamatergic neurotransmission in the ventral tegmental area. *Eur. J. Neurosci.* *13*, 819–828.
- Leutgeb, S., Leutgeb, J.K., Barnes, C.A., Moser, E.I., McNaughton, B.L., and Moser, M.-B. (2005). Independent codes for spatial and episodic memory in hippocampal neuronal ensembles. *Science* *309*, 619–623.
- Lever, C., Wills, T., Cacucci, F., Burgess, N., and O’Keefe, J. (2002). Long-term plasticity in hippocampal place-cell representation of environmental geometry. *Nature* *416*, 90–94.
- Li, B., and Freeman, R.D. (2007). High-Resolution Neurometabolic Coupling in the Lateral Geniculate Nucleus. *J. Neurosci.* *27*, 10223–10229.



- 
- Li, L., Miller, E.K., and Desimone, R. (1993a). The representation of stimulus familiarity in anterior inferior temporal cortex. *J. Neurophysiol.*
- Li, L., Miller, E.K., and Desimone, R. (1993b). The representation of stimulus familiarity in anterior inferior temporal cortex. *J. Neurophysiol.* *69*, 1918–1929.
- Li, S., Cullen, W.K., Anwyl, R., and Rowan, M.J. (2003). Dopamine-dependent facilitation of LTP induction in hippocampal CA1 by exposure to spatial novelty. *Nat. Neurosci.* *6*, 526–531.
- Lisman, J.E., and Grace, A.A. (2005). The hippocampal-VTA loop: controlling the entry of information into long-term memory. *Neuron* *46*, 703–713.
- Liu, Y., Murray, S.O., and Jagadeesh, B. (2009). Time course and stimulus dependence of repetition-induced response suppression in inferotemporal cortex. *J. Neurophysiol.* *101*, 418–436.
- Logothetis, N.K. (2003). The Underpinnings of the BOLD Functional Magnetic Resonance Imaging Signal. *J. Neurosci.* *23*, 3963–3971.
- Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature* *412*, 150–157.
- MacDonald, C.J., Lepage, K.Q., Eden, U.T., and Eichenbaum, H. (2011). Hippocampal “time cells” bridge the gap in memory for discontinuous events. *Neuron* *71*, 737–749.
- Madden, G.J., Petry, N.M., Badger, G.J., and Bickel, W.K. (1997). Impulsive and self-control choices in opioid-dependent patients and non-drug-using control participants: drug and monetary rewards. *Exp. Clin. Psychopharmacol.* *5*, 256–262.
- Magno, E., and Allan, K. (2007). Self-reference during explicit memory retrieval an event-related potential analysis. *Psychol. Sci.* *18*, 672–677.
- Malach, R. (2012). Targeting the functional properties of cortical neurons using fMR-adaptation. *Neuroimage* *62*, 1163–1169.
- Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* *503*, 78–84.
- De Martino, B., Fleming, S.M., Garrett, N., and Dolan, R.J. (2013). Confidence in value-based choice. *Nat. Neurosci.* *16*, 105–110.
- Mathis, A., Herz, A., and Stemmler, M. (2012). Optimal population codes for space: grid cells outperform place cells. *Neural Comput.*
- McClure, S.M., Berns, G.S., and Montague, P.R. (2003). Temporal Prediction Errors in a Passive Learning Task Activate Human Striatum. *Neuron* *38*, 339–346.
- McClure, S.M., Li, J., Tomlin, D., Cypert, K.S., Montague, L.M., and Montague, P.R. (2004). Neural correlates of behavioral preference for culturally familiar drinks. *Neuron* *44*, 379–387.
- McKenzie, S., Frank, A., Kinsky, N., and Porter, B. (2014). Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. *Neuron*.
- McMahon, D.B.T., and Olson, C.R. (2007). Repetition Suppression in Monkey Inferotemporal Cortex: Relation to Behavioral Priming. *J. Neurophysiol.* *97*, 3532–3543.
- McNaughton, B.L., Battaglia, F.P., Jensen, O., Moser, E.I., and Moser, M.-B. (2006). Path integration and the neural basis of the “cognitive map”. *Nat. Rev. Neurosci.* *7*, 663–678.

- 
- Mehta, M.R., Lee, A.K., and Wilson, M.A. (2002). Role of experience and oscillations in transforming a rate code into a temporal code. *Nature* *417*, 741–746.
- Meyer, T., and Olson, C.R. (2011). Statistical learning of visual transitions in monkey inferotemporal cortex. *Proc. Natl. Acad. Sci.* *108*, 19401–19406.
- Michalareas, G., Vezoli, J., van Pelt, S., Schoffelen, J.-M., Kennedy, H., and Fries, P. (2016). Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas. *Neuron* *89*, 384–397.
- Miller, E.K., and Desimone, R. (1994). Parallel neuronal mechanisms for short-term memory. *Science* (80-. ). *263*, 520–522.
- Miller, E.K., Li, L., and Desimone, R. (1991). A neural mechanism for working and recognition memory in inferior temporal cortex. *Science* *254*, 1377–1379.
- Miller, E.K., Erickson, C.A., and Desimone, R. (1996). Neural Mechanisms of Visual Working Memory in Prefrontal Cortex of the Macaque. *J. Neurosci.* *16*, 5154–5167.
- Mitchell, J.P. (2009). Inferences about mental states. *364*, 1309–1316.
- Mitchell, J.P., Macrae, C.N., and Banaji, M.R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron* *50*, 655–663.
- Mitchell, J.P., Schirmer, J., Ames, D.L., and Gilbert, D.T. (2010). Medial prefrontal cortex predicts intertemporal choice. *J. Cogn. Neurosci.* *23*, 857–866.
- Mittelstaedt, M.-L., and Mittelstaedt, H. (1980). Homing by path integration in a mammal. *Naturwissenschaften* *67*, 566–567.
- Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* *335*, 817–820.
- Mizumori, S., and Williams, J. (1993). Directionally selective mnemonic properties of neurons in the lateral dorsal nucleus of the thalamus of rats. *J. Neurosci.* *13*, 4015–4028.
- Morgan, L.K., Macevoy, S.P., Aguirre, G.K., and Epstein, R.A. (2011). Distances between real-world locations are represented in the human hippocampus. *J. Neurosci.* *31*, 1238–1245.
- Morris, J.S., Frith, C.D., Perrett, D.I., Rowland, D., Young, A.W., Calder, A.J., and Dolan, R.J. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature* *383*, 812–815.
- Morris, R.G.M., Moser, E.I., Riedel, G., Martin, S.J., Sandin, J., Day, M., and O’Carroll, C. (2003). Elements of a neurobiological theory of the hippocampus: the role of activity-dependent synaptic plasticity in memory. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* *358*, 773–786.
- Moser, E.I., Roudi, Y., Witter, M.P., Kentros, C., Bonhoeffer, T., and Moser, M.-B. (2014). Grid cells and cortical representation. *Nat. Rev. Neurosci.* *15*, 466–481.
- Muessig, L., Hauser, J., Wills, T.J., and Cacucci, F. (2015). A Developmental Switch in Place Cell Accuracy Coincides with Grid Cell Maturation. *Neuron* *86*, 1167–1173.
- Muir, G.M., and Taube, J.S. (2004). Head direction cell activity and behavior in a navigation task requiring a cognitive mapping strategy. *Behav. Brain Res.* *153*, 249–253.
- Muller, R., and Kubie, J. (1987). The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *J. Neurosci.* *7*, 1935–1950.
- Muller, T., Baram, A., and Behrens, T.E.J. Intuitive planning: global navigation through

---

cognitive maps based on grid-like codes.

Myerson, J., and Green, L. (1995). Discounting of delayed rewards: Models of individual choice. *J. Exp. Anal. Behav.* *64*, 263–276.

Nääätänen, R., Paavilainen, P., Alho, K., Reinikainen, K., and Sams, M. (1989). Do event-related potentials reveal the mechanism of the auditory sensory memory in the human brain? *Neurosci. Lett.* *98*, 217–221.

Nabavi, S., Fox, R., Proulx, C.D., Lin, J.Y., Tsien, R.Y., and Malinow, R. (2014). Engineering a memory with LTD and LTP. *Nature* *511*, 348–352.

Nass, C., and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *J. Soc. Issues* *56*, 81–103.

Naya, Y., and Suzuki, W.A. (2011). Integrating what and when across the primate medial temporal lobe. *Science* *333*, 773–776.

Neves, G., Cooke, S.F., and Bliss, T.V.P. (2008). Synaptic plasticity, memory and the hippocampus: a neural network approach to causality. *Nat. Rev. Neurosci.* *9*, 65–75.

Nicolle, A., Klein-Flügge, M.C., Hunt, L.T., Vlaev, I., Dolan, R.J., and Behrens, T.E.J. (2012). An agent independent axis for executed and modeled choice in medial prefrontal cortex. *Neuron* *75*, 1114–1121.

Niv, Y., Edlund, J.A., Dayan, P., and O’Doherty, J.P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *J. Neurosci.* *32*, 551–562.

Norman, K.A., and O’Reilly, R.C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychol. Rev.*

Norman, K.A., Polyn, S.M., Detre, G.J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* *10*, 424–430.

O’Doherty, J.P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Curr. Opin. Neurobiol.* *14*, 769–776.

O’Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D., and Dolan, R. (2003a). Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia* *41*, 147–155.

O’Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R.J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* (80-. ). *304*, 452–454.

O’Doherty, J.P., Dayan, P., Friston, K., Critchley, H., and Dolan, R.J. (2003b). Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron* *38*, 329–337.

O’Keefe, J. (1976). Place units in the hippocampus of the freely moving rat. *Exp. Neurol.* *51*, 78–109.

O’Keefe, J., and Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. *Nature* *381*, 425–428.

O’Keefe, J., and Burgess, N. (2005). Dual phase and rate coding in hippocampal place cells: theoretical significance and relationship to entorhinal grid cells. *Hippocampus* *15*, 853–866.

O’Keefe, J., and Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* *34*, 171–175.

- 
- O'Keefe, J., and Nadel, L. (1978). *The hippocampus as a cognitive map* (Oxford: Clarendon Press).
- O'Keefe, J., and Recce, M.L. (1993). Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* *3*, 317–330.
- O'Neill, J., Senior, T.J., Allen, K., Huxter, J.R., and Csicsvari, J. (2008). Reactivation of experience-dependent cell assembly patterns in the hippocampus. *Nat. Neurosci.* *11*, 209–215.
- O'Reilly, R.C., and McClelland, J.L. (1994). Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus* *4*, 661–682.
- O'Reilly, J.X., Schüffelgen, U., Cuell, S.F., Behrens, T.E.J., Mars, R.B., and Rushworth, M.F.S. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proc. Natl. Acad. Sci. U. S. A.* *110*, E3660–E3669.
- Obermayer, K., and Blasdel, G.G. (1993). Geometry of orientation and ocular dominance columns in monkey striate cortex. *J. Neurosci.* *13*, 4114–4129.
- Ogawa, S., Lee, T.M., Kay, A.R., and Tank, D.W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci.* *87*, 9868–9872.
- Ohmura, Y., Takahashi, T., Kitamura, N., and Wehr, P. (2006). Three-month stability of delay and probability discounting measures. *Exp. Clin. Psychopharmacol.* *14*, 318–328.
- Okun, M., and Lampl, I. (2008). Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nat. Neurosci.* *11*, 535–537.
- Ólafsdóttir, H.F., Barry, C., Saleem, A.B., Hassabis, D., and Spiers, H.J. (2015). Hippocampal place cells construct reward related sequences through unexplored space. *Elife* *4*, e06063.
- Ouden, H.E.M. den, Friston, K.J., Daw, N.D., McIntosh, A.R., and Stephan, K.E. (2009). A Dual Role for Prediction Error in Associative Learning. *Cereb. Cortex* *19*, 1175–1185.
- den Ouden, H.E.M., Daunizeau, J., Roiser, J., Friston, K.J., and Stephan, K.E. (2010). Striatal prediction error modulates cortical coupling. *J. Neurosci.* *30*, 3210–3219.
- Packard, M.G., and Knowlton, B.J. (2002). Learning and memory functions of the Basal Ganglia. *Annu. Rev. Neurosci.* *25*, 563–593.
- Park, S.Q., Kahnt, T., Rieskamp, J., and Heekeren, H.R. (2011). Neurobiology of value integration: when value impacts valuation. *J. Neurosci.* *31*, 9307–9314.
- Pastalkova, E., Serrano, P., Pinkhasova, D., Wallace, E., Fenton, A.A., and Sacktor, T.C. (2006). Storage of spatial information by the maintenance mechanism of LTP. *Science* *313*, 1141–1144.
- Pastalkova, E., Itskov, V., Amarasingham, A., and Buzsáki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science* *321*, 1322–1327.
- Pauling, L., and Coryell, C.D. (1936). The Magnetic Properties and Structure of Hemoglobin, Oxyhemoglobin and Carbonmonoxyhemoglobin. *Proc. Natl. Acad. Sci. U. S. A.* *22*, 210–216.
- Pavlov, I.P. (1849-1936) (1927). *Conditioned reflexes : an investigation of the physiological activity of the cerebral cortex.*
- Paz-Y-Miño C, G., Bond, A.B., Kamil, A.C., and Balda, R.P. (2004). Pinyon jays use transitive inference to predict social dominance. *Nature* *430*, 778–781.

- 
- Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., and Frith, C.D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* *442*, 1042–1045.
- Peters, J., and Büchel, C. (2010). Episodic future thinking reduces reward delay discounting through an enhancement of prefrontal-mediotemporal interactions. *Neuron* *66*, 138–148.
- Pfeiffer, B.E., and Foster, D.J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* *497*, 74–79.
- Piazza, M., Pinel, P., Le Bihan, D., and Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron* *53*, 293–305.
- Pine, A., Seymour, B., Roiser, J.P., Bossaerts, P., Friston, K.J., Curran, H.V., and Dolan, R.J. (2009). Encoding of marginal utility across time in the human brain. *J. Neurosci. Off. J. Soc. Neurosci.* *29*, 9575–9581.
- Plassmann, H., O’Doherty, J., and Rangel, A. (2007). Orbitofrontal Cortex Encodes Willingness to Pay in Everyday Economic Transactions. *J. Neurosci.* *27*, 9984–9988.
- Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nat. Rev. Neurosci.* *1*, 125–132.
- Premack, D., and Woodruff, G. (1978). Does the Chimpanzee Have a Theory of Mind? *Behav. Brain Sci.* *1*, 515–526.
- Preston, A.R., Shrager, Y., Dudukovic, N.M., and Gabrieli, J.D.E. (2004). Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus* *14*, 148–152.
- Purcell, E.M., Torrey, H.C., and Pound, R. V. (1946). Resonance Absorption by Nuclear Magnetic Moments in a Solid. *Phys. Rev.* *69*, 37–38.
- Quian Quiroga, R., and Panzeri, S. (2009). Extracting information from neuronal populations: information theory and decoding approaches. *Nat. Rev. Neurosci.* *10*, 173–185.
- Quirk, G., Muller, R., and Kubie, J. (1990). The firing of hippocampal place cells in the dark depends on the rat’s recent experience. *J. Neurosci.* *10*, 2008–2017.
- Rachlin, H., Raineri, A., and Cross, D. (1991). Subjective probability and delay. *J. Exp. Anal. Behav.* *55*, 233–244.
- Rao, R.P.N., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* *2*, 79–87.
- Redish, A.D., Jensen, S., Johnson, A., and Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychol. Rev.* *114*, 784–805.
- Rees, G., Friston, K., and Koch, C. (2000). A direct quantitative relationship between the functional properties of human and macaque V5. *Nat. Neurosci.* *3*, 716–723.
- Rescorla, R., and Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Class. Cond. Curr. Res. Theory.*
- Reynolds, J.N.J., and Wickens, J.R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks* *15*, 507–521.
- Riches, I.P., Wilson, F.A., and Brown, M.W. (1991). The effects of visual stimulation and memory on neurons of the hippocampal formation and the neighboring parahippocampal gyrus and inferior temporal cortex of the primate. *J. Neurosci.* *11*, 1763–1779.

- 
- Rigotti, M., Barak, O., Warden, M.R., Wang, X.-J., Daw, N.D., Miller, E.K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature* *497*, 585–590.
- Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cogn. Brain Res.* *3*, 131–141.
- Rizzolatti, G., Fogassi, L., and Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nat. Rev. Neurosci.* *2*, 661–670.
- Robbins, T.W., Gillan, C.M., Smith, D.G., Wit, S. de, and Ersche, K.D. (2012). Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry. *Trends Cogn. Sci.* *16*, 81–91.
- Roediger, H.L.I.H., and McDermott, K.B. (1993). Implicit memory in normal human subjects. *Handb. Neuropsychol.* *8*.
- Roesch, M.R., Taylor, A.R., and Schoenbaum, G. (2006). Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation. *Neuron* *51*, 509–520.
- Rolls, E.T. (1999). The functions of the orbitofrontal cortex. *Neurocase* *5*, 301–312.
- Rolls, E.T., Sienkiewicz, Z.J., and Yaxley, S. (1989). Hunger Modulates the Responses to Gustatory Stimuli of Single Neurons in the Caudolateral Orbitofrontal Cortex of the Macaque Monkey. *Eur. J. Neurosci.* *1*, 53–60.
- Rolls, E.T., Stringer, S.M., and Elliot, T. (2006). Entorhinal cortex grid cells can map to hippocampal place cells by competitive learning. *Network* *17*, 447–465.
- Rossion, B., and Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart’s object pictorial set: the role of surface detail in basic-level object recognition. *Perception* *33*, 217–236.
- Ruff, C.C., Ugazio, G., and Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science* (80-. ). *342*, 482–484.
- Rutledge, R.B., Lazzaro, S.C., Lau, B., Myers, C.E., Gluck, M.A., and Glimcher, P.W. (2009). Dopaminergic drugs modulate learning rates and perseveration in Parkinson’s patients in a dynamic foraging task. *J. Neurosci.* *29*, 15104–15114.
- Sanchez-Vives, M. V., and McCormick, D.A. (2000). Cellular and network mechanisms of rhythmic recurrent activity in neocortex. *Nat. Neurosci.* *3*, 1027–1034.
- Sapountzis, P., Schluppeck, D., Bowtell, R., and Peirce, J.W. (2010). A comparison of fMRI adaptation and multivariate pattern classification analysis in visual cortex. *Neuroimage* *49*, 1632–1640.
- Sargolini, F., Fyhn, M., Hafting, T., McNaughton, B.L., Witter, M.P., Moser, M.-B., and Moser, E.I. (2006). Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science* *312*, 758–762.
- Sawamura, H., Orban, G.A., and Vogels, R. (2006). Selectivity of Neuronal Adaptation Does Not Match Response Selectivity: A Single-Cell Study of the fMRI Adaptation Paradigm. *Neuron* *49*, 307–318.
- Saxe, R. (2006). Uniquely human social cognition. *Curr. Opin. Neurobiol.* *16*, 235–239.
- Schacter, D.L., and Buckner, R.L. (1998). Priming and the Brain. *Neuron* *20*, 185–195.
- Schacter, D.L., and Church, B.A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *J. Exp. Psychol. Learn. Mem. Cogn.* *18*, 915–930.

- 
- Schacter, D.L.D., Addis, D.R.D., Hassabis, D., Martin, V.C., Spreng, R.N., and Szpunar, K.K. (2012). The future of memory: remembering, imagining, and the brain. *Neuron* *76*, 677–694.
- Schapiro, A.C., Kustner, L.V., and Turk-Browne, N.B. (2012). Shaping of Object Representations in the Human Medial Temporal Lobe Based on Temporal Regularities. *Curr. Biol.* *22*, 1622–1627.
- Schapiro, A.C., Rogers, T.T., Cordova, N.I., Turk-Browne, N.B., and Botvinick, M.M. (2013). Neural representations of events arise from temporal community structure. *Nat. Neurosci.* *16*, 486–492.
- Schiller, D., Eichenbaum, H., Buffalo, E.A., Davachi, L., Foster, D.J., Leutgeb, S., and Ranganath, C. (2015). Memory and Space: Towards an Understanding of the Cognitive Map. *J. Neurosci.* *35*, 13904–13911.
- Schlichting, M.L., Mumford, J.A., and Preston, A.R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nat. Commun.* *6*, 8151.
- Schonberg, T., O’Doherty, J.P., Joel, D., Inzelberg, R., Segev, Y., and Daw, N.D. (2010). Selective impairment of prediction error signaling in human dorsolateral but not ventral striatum in Parkinson’s disease patients: evidence from a model-based fMRI study. *Neuroimage* *49*, 772–781.
- Schönberg, T., Daw, N.D., Joel, D., and O’Doherty, J.P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *J. Neurosci.* *27*, 12860–12867.
- Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. *Science* *275*, 1593–1599.
- Scoville, W.B., and Milner, B. (1957). Loss of Recent Memory After Bilateral. *J. Neuropsychiatry Clin. Neurosci.* *12*, 103–113.
- Shestakova, A., Rieskamp, J., Tugin, S., Ossadtchi, A., Krutitskaya, J., and Klucharev, V. (2013). Electrophysiological precursors of social conformity. *Soc. Cogn. Affect. Neurosci.* *8*, 756–763.
- Sheth, S.A., Nemoto, M., Guiou, M., Walker, M., Pouratian, N., and Toga, A.W. (2004). Linear and Nonlinear Relationships between Neuronal Activity, Oxygen Metabolism, and Hemodynamic Responses. *Neuron* *42*, 347–355.
- Singer, T., Seymour, B., O’Doherty, J., Kaube, H., Dolan, R.J., and Frith, C.D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science* *303*, 1157–1162.
- Skaggs, W.E., and McNaughton, B.L. (1996). Replay of Neuronal Firing Sequences in Rat Hippocampus During Sleep Following Spatial Experience. *Science* (80-. ). *271*, 1870–1873.
- Skaggs, W.E., McNaughton, B.L., Wilson, M.A., and Barnes, C.A. (1996). Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus* *6*, 149–172.
- Skinner, B. (1938). The behavior of organisms: an experimental analysis.
- Sobotka, S., and Ringo, J.L. (1994). Stimulus specific adaptation in excited but not in inhibited cells in inferotemporal cortex of Macaque. *Brain Res.* *646*, 95–99.
- Solstad, T., Moser, E.I., and Einevoll, G.T. (2006). From grid cells to place cells: a

- 
- mathematical model. *Hippocampus* *16*, 1026–1031.
- Solstad, T., Boccarda, C.N., Kropff, E., Moser, M.-B., and Moser, E.I. (2008). Representation of geometric borders in the entorhinal cortex. *Science* *322*, 1865–1868.
- Spiers, H.J., and Maguire, E.A. (2007). A navigational guidance system in the human brain. *Hippocampus* *17*, 618–626.
- Stachenfeld, K.L., Botvinick, M., and Gershman, S.J. (2014). Design Principles of the Hippocampal Cognitive Map. In *Advances in Neural Information Processing Systems*, pp. 2528–2536.
- Stefanics, G., Kimura, M., and Czigler, I. (2011). Visual mismatch negativity reveals automatic detection of sequential regularity violation. *Front. Hum. Neurosci.* *5*, 46.
- Steinberg, E.E., Keiflin, R., Boivin, J.R., Witten, I.B., Deisseroth, K., and Janak, P.H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* *16*, 966–973.
- Stemmler, M., Mathis, A., and Herz, A.V.M. (2015). Connecting multiple spatial scales to decode the population activity of grid cells. *Sci. Adv.* *1*, e1500816–e1500816.
- Stensola, H., Stensola, T., Solstad, T., Frøland, K., Moser, M.-B., and Moser, E.I. (2012). The entorhinal grid map is discretized. *Nature* *492*, 72–78.
- Steyvers, M., Griffiths, T.L., and Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends Cogn. Sci.* *10*, 327–334.
- Strange, B.A., Duggins, A., Penny, W., Dolan, R.J., and Friston, K.J. (2005a). Information theory, novelty and hippocampal responses: unpredicted or unpredictable? *Neural Networks* *18*, 225–230.
- Strange, B.A., Hurlemann, R., Duggins, A., Heinze, H.-J., and Dolan, R.J. (2005b). Dissociating intentional learning from relative novelty responses in the medial temporal lobe. *Neuroimage* *25*, 51–62.
- Strange, B.A., Witter, M.P., Lein, E.S., and Moser, E.I. (2014). Functional organization of the hippocampal longitudinal axis. *Nat. Rev. Neurosci.* *15*, 655–669.
- Summerfield, C., Trittschuh, E.H., Monti, J.M., Mesulam, M.-M., and Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nat. Neurosci.* *11*, 1004–1006.
- Summerfield, C., Wyart, V., Mareike Johnen, V., and de Gardelle, V. (2011). Human scalp electroencephalography reveals that repetition suppression varies with expectation. *Front. Hum. Neurosci.* *5*, 67.
- Sutton, R.S. (1988). Learning to predict by the methods of temporal differences. *Mach. Learn.* *3*, 9–44.
- Sutton, R.S., and Barto, A.G. (1990). Time-derivative models of Pavlovian reinforcement. In *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, pp. 497–537.
- Suzuki, S., Harasawa, N., Ueno, K., Gardner, J.L., Ichinohe, N., Haruno, M., Cheng, K., and Nakahara, H. (2012). Learning to simulate others' decisions. *Neuron* *74*, 1125–1137.
- Swanson, L.W. (1981). A direct projection from Ammon's horn to prefrontal cortex in the rat. *Brain Res.* *217*, 150–154.
- Tamir, D.I., and Mitchell, J.P. (2011). The default network distinguishes construals of



- 
- proximal versus distal events. *J. Cogn. Neurosci.* *23*, 2945–2955.
- Tanaka, K.Z., Pevzner, A., Hamidi, A.B., Nakazawa, Y., Graham, J., and Wiltgen, B.J. (2014). Cortical Representations Are Reinstated by the Hippocampus during Memory Retrieval. *Neuron* *84*, 347–354.
- Tanaka, S.C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., and Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat. Neurosci.* *7*, 887–893.
- Taube, J., Muller, R., and Ranck, J.J. (1990). Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *J. Neurosci.* *10*, 420–435.
- Tavares, R.M.M., Mendelsohn, A., Grossman, Y., Williams, C.H.H., Shapiro, M., Trope, Y., and Schiller, D. (2015). A Map for Social Navigation in the Human Brain. *Neuron* *87*, 231–243.
- Terrazas, A., Krause, M., Lipa, P., Gothard, K.M., Barnes, C.A., and McNaughton, B.L. (2005). Self-motion and the hippocampal spatial metric. *J. Neurosci.* *25*, 8085–8096.
- Tervo, D.G.R., Tenenbaum, J.B., and Gershman, S.J. (2016). Toward the neural implementation of structure learning. *Curr. Opin. Neurobiol.* *37*, 99–105.
- Thomsen, K., Offenhauser, N., and Lauritzen, M. (2004). Principal neuron spiking: neither necessary nor sufficient for cerebral blood flow in rat cerebellum. *J. Physiol.* *560*, 181–189.
- Thorndike, E. (1898). *Animal intelligence: An experimental study of the associative processes in animals.* Psychol. Rev. Monogr. ....
- Thorndike, E. (1927). The law of effect. *Am. J. Psychol.*
- Todorovic, A., and Lange, F.P. de (2012). Repetition Suppression and Expectation Suppression Are Dissociable in Time in Early Auditory Evoked Fields. *J. Neurosci.* *32*, 13389–13395.
- Tolman, E.C. (1948). Cognitive maps in rats and men. *Psychol. Rev.* *55*, 189–208.
- Tolman, E.C., and Honzik, C.H. (1930). Introduction and removal of reward, and maze performance in rats. *J. Psychol.* *4*, 241–256.
- Tomlin, D., Nedic, A., Prentice, D.A., Holmes, P., and Cohen, J.D. (2013). The neural substrates of social influence on decision making. *PLoS One* *8*, e52630.
- Treves, A., and Rolls, E.T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus* *4*, 374–391.
- Trope, Y., and Liberman, N. (2010). Construal-level theory of psychological distance. *Psychol. Rev.* *117*, 440–463.
- Tse, D., Langston, R.F., Kakeyama, M., Bethus, I., Spooner, P.A., Wood, E.R., Witter, M.P., and Morris, R.G.M. (2007). Schemas and Memory Consolidation. *Science* (80-. ). *316*, 76–82.
- Tulving, E., and Schacter, D.L. (1990). Priming and human memory systems. *Science* (80-. ). *247*, 301–306.
- Valentini, E., Torta, D.M.E., Mouraux, A., and Iannetti, G.D. (2011). Dishabituation of Laser-evoked EEG Responses: Dissecting the Effect of Certain and Uncertain Changes in Stimulus Modality. *J. Cogn. Neurosci.* *23*, 2822–2837.

- 
- Voon, V., Pessiglione, M., Brezing, C., Gallea, C., Fernandez, H.H., Dolan, R.J., and Hallett, M. (2010). Mechanisms underlying dopamine-mediated reward bias in compulsive behaviors. *Neuron* *65*, 135–142.
- Vuilleumier, P., Henson, R.N., Driver, J., and Dolan, R.J. (2002). Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. *Nat. Neurosci.* *5*, 491–499.
- Wager, T.D., Davidson, M.L., Hughes, B.L., Lindquist, M.A., and Ochsner, K.N. (2008). Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron* *59*, 1037–1050.
- Walton, M.E., Behrens, T.E.J., Buckley, M.J., Rudebeck, P.H., and 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.
- Wang, G., Tanaka, K., and Tanifuji, M. (1996). Optical imaging of functional organization in the monkey inferotemporal cortex. *Science* *272*, 1665–1668.
- Wang, Q., Luo, S., Monterosso, J., Zhang, J., Fang, X., Dong, Q., and Xue, G. (2014). Distributed value representation in the medial prefrontal cortex during intertemporal choices. *J. Neurosci.* *34*, 7522–7530.
- Watson, J.B. (1913). Psychology as the behaviorist views it. *Psychol. Rev.* *20*, 158–177.
- Weber, E. (1834). *De Pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae...*
- Whishaw, I.Q., and Brooks, B.L. (1999). Calibrating space: exploration is important for allothetic and idiothetic navigation. *Hippocampus* *9*, 659–667.
- Whitlock, J.R., Heynen, A.J., Shuler, M.G., and Bear, M.F. (2006). Learning induces long-term potentiation in the hippocampus. *Science* *313*, 1093–1097.
- Wickens, J.R., Begg, A.J., and Arbuthnott, G.W. (1996). Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex *In vitro*. *Neuroscience* *70*, 1–5.
- Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V., and Rizzolatti, G. (2003). Both of us disgusted in my insula. *Neuron* *40*, 655–664.
- Wiestler, T., McGonigle, D.J., and Diedrichsen, J. (2011). Integration of sensory and motor representations of single fingers in the human cerebellum. *J. Neurophysiol.* *105*, 3042–3053.
- Wiggs, C.L., and Martin, A. (1998). Properties and mechanisms of perceptual priming. *Curr. Opin. Neurobiol.* *8*, 227–233.
- Wills, T.J., Lever, C., Cacucci, F., Burgess, N., and O’Keefe, J. (2005). Attractor dynamics in the hippocampal representation of the local environment. *Science* *308*, 873–876.
- Wills, T.J., Cacucci, F., Burgess, N., and O’Keefe, J. (2010). Development of the hippocampal cognitive map in preweanling rats. *Science* *328*, 1573–1576.
- Wilson, M., and McNaughton, B. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science* (80-. ). *265*, 676–679.
- Wilson, M.A., and McNaughton, B.L. (1993). Dynamics of the hippocampal ensemble code for space. *Science* (80-. ). *261*, 1055–1058.
- Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., and Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron* *81*, 267–279.

- 
- Wimmer, G.E., and Shohamy, D. (2012). Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science* 338, 270–273.
- Winstanley, C.A., Eagle, D.M., and Robbins, T.W. (2006). Behavioral models of impulsivity in relation to ADHD: Translation between clinical and preclinical studies. *Clin. Psychol. Rev.* 26, 379–395.
- Wood, E.R., Dudchenko, P.A., Robitsek, R.J., and Eichenbaum, H. (2000). Hippocampal Neurons Encode Information about Different Types of Memory Episodes Occurring in the Same Location. *Neuron* 27, 623–633.
- Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J.A., and Poldrack, R.A. (2010). Greater neural pattern similarity across repetitions is associated with better memory. *Science* 330, 97–101.
- Yartsev, M.M., and Ulanovsky, N. (2013). Representation of three-dimensional space in the hippocampus of flying bats. *Science* 340, 367–372.
- Yin, H.H., Knowlton, B.J., and 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., and Balleine, B.W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *Eur. J. Neurosci.* 22, 513–523.
- Yoshida, W., Seymour, B., Friston, K.J., and Dolan, R.J. (2010). Neural mechanisms of belief inference during cooperative games. *J. Neurosci.* 30, 10744–10751.
- Zaghloul, K.A., Blanco, J.A., Weidemann, C.T., McGill, K., Jaggi, J.L., Baltuch, G.H., and Kahana, M.J. (2009). Human substantia nigra neurons encode unexpected financial rewards. *Science* 323, 1496–1499.
- Zaki, J., Schirmer, J., and Mitchell, J.P. (2011). Social influence modulates the neural computation of value. *Psychol. Sci.* 22, 894–900.
- Zhang, K., Ginzburg, I., McNaughton, B.L., and Sejnowski, T.J. (1998). Interpreting Neuronal Population Activity by Reconstruction: Unified Framework With Application to Hippocampal Place Cells. *J. Neurophysiol.* 79, 1017–1044.