Predictive maps in rats and humans for spatial navigation

William John de Cothi

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

CoMPLEX
University College London

May 27, 2020

I, William John de Cothi, 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 work.

Abstract

The ability to navigate space is an essential part of mammalian life. Over the last 50 years, much research has investigated on how the mammalian brain represents space in the activity of populations of neurons, particularly focussing upon cells in the hippocampal formulation. But how does the brain integrate these representations to guide flexible and efficient navigational decision making? A useful way to approach this question is from the field of reinforcement learning, which seeks to address how an agent should act in its environment in order to maximise some form of reward signal. Typically solutions to a reinforcement learning problem are split into a dichotomy of model-free and model-based approaches. Here we investigate the biological validity of an intermediary approach called the successor representation, which works by forming a predictive map of the environment. First, we compare these three reinforcement learning methods to rat and human behaviour on a transition revaluation spatial navigation task, and show that the biological behaviour is most similar to that of a successor representation agent. Then we propose a neurally plausible implementation of the successor representation, based upon a set of known neurobiological features - boundary vector cells. We show that the place and grid cells generated using this model provide a good account of biological data for a variety of environmental manipulations, including dimensional stretches, barrier insertions, and the influence of environmental geometry on the hippocampal representation of space.

Impact Statement

The speculative benefits inside academia might be to the discipline and future of scholarship, research methods or methodology, the curriculum; they might be within neuroscience and potentially within other research areas.

The speculative benefits outside academia might be to commercial activity, social enterprise, professional practice, clinical use, public health, public policy design, public service delivery, laws, public discourse, culture, the quality of the environment or quality of life.

There might not even be any benefits at all.

Statement of Contribution

The work presented in this thesis would not have been possible without the contributions of many brilliant people, to whom I am very grateful.

The rodent data used in Chapter 2 was collected by Eva-Maria Griesbauer, Carole Ghanamé and Nils Nyberg under the supervision of Hugo Spiers. The rodent maze itself was built by Célia Lacaux and Charles Middleton under the supervision of Hugo Spiers. The human data used in Chapter 2 was collected by Lydia Fletcher and the author, in a virtual environment designed and built by the author and under the supervision of Hugo Spiers. All subsequent analysis and simulations of reinforcement learning agents was carried out by the author under the supervision of Caswell Barry and Hugo Spiers.

All modelling, accompanying simulations and analysis presented in Chapter 3 was carried out by the author and under the supervision of Caswell Barry.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Neural basis of spatial navigation

1.1.1 What is spatial navigation?

For all mobile organisms, including mammals such as ourselves, knowing where we are and how to get to resources such as food and safety is a crucial part of everyday life. In order to locate yourself in a space, you can only define your position with respect to something else; be it the walls of a room, a set of landmarks in the local area, or even a previously occupied position in space. How the brain robustly integrates these points of reference in order to decide upon a sequence of actions that lead to a specific goal location is what we refer to as the neural basis of spatial navigation.

Behavioural evidence that small mammals, such as rodents, form a cognitive representation of their environment has been around since Tolman and Honzik (1930). They demonstrated that rats were better at finding a food reward on a maze if they were first familiarised to it in the absence of the reward (figure 1.1A). This was compared to other rats that had the food reward present the whole time. This phenomenon is described as *latent learning* and suggests that the rats develop a better

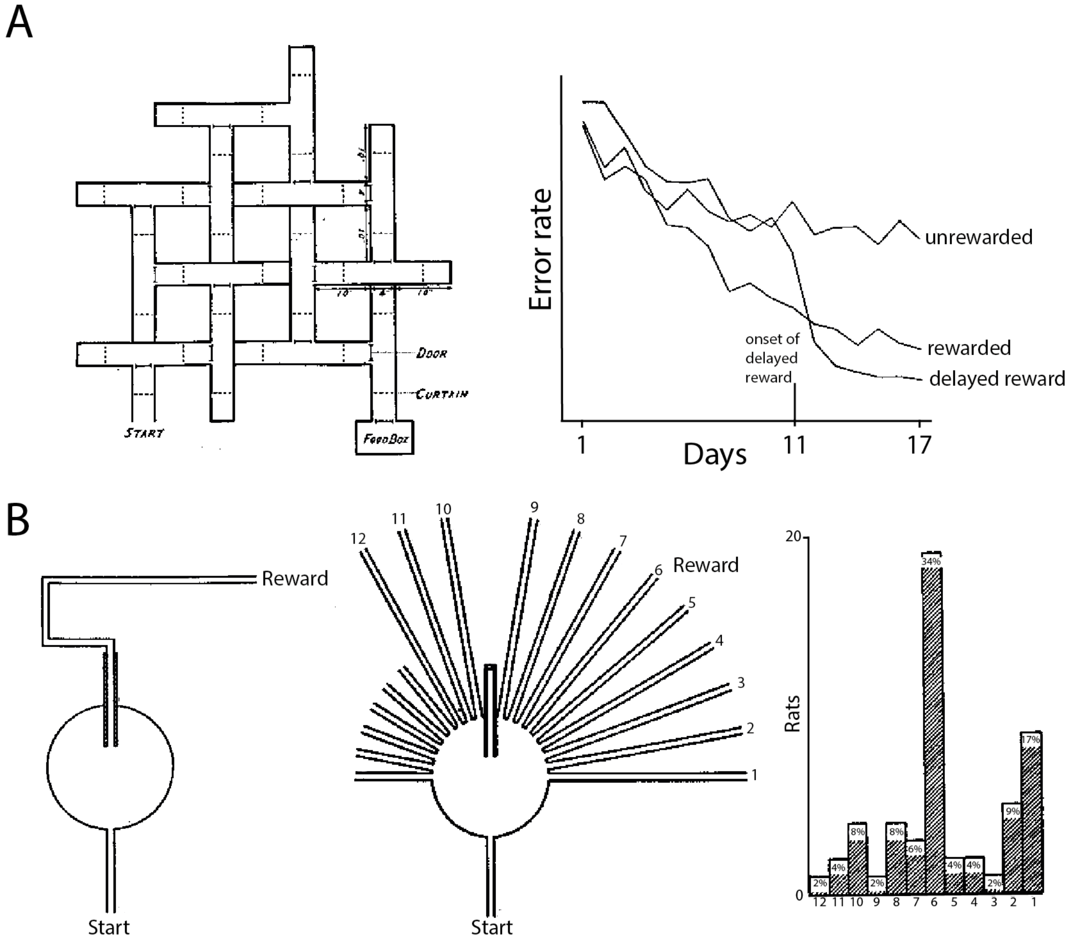


Figure 1.1: (A) Rats were made to navigate a maze from ‘start’ to ‘food box’ (left). Presence of food reward in the ‘food box’ led to improved rat performance on the maze. However, rats that were introduced to this food reward after 10 days of exposure to the maze without reward formed a cognitive map and were subsequently better than rats that had the food reward present the entire time (right). (B) Rats were trained to navigate a circuitous route to a food reward (left). Next, this route was blocked and the rats were given a range of alternative options (middle). The majority of rats chose the new route that took them directly to the reward, suggesting they are able to compute a vector to the goal location (right). Figures adapted from Tolman (1948).

internal representation of the maze, or *cognitive map*, during the pre-reward familiarisation phase. It is argued that this subsequently allows them to choose better actions, such as turning down less dead ends. The existence of a cognitive map was further evidenced by rats ability to take optimal novel shortcuts when a previously learnt route was blocked (Tolman, 1948). After rats had been trained to navigate via a circuitous route to a food reward, this route was blocked and they were presented with a range of alternatives - all of which traversed through previously unvisited space (figure 1.1B). The majority of the rats chose the new route that minimised their distance and direction to the goal, suggesting not only the existence of a cognitive map, but that this map is able to facilitate vector-based navigation strategies.

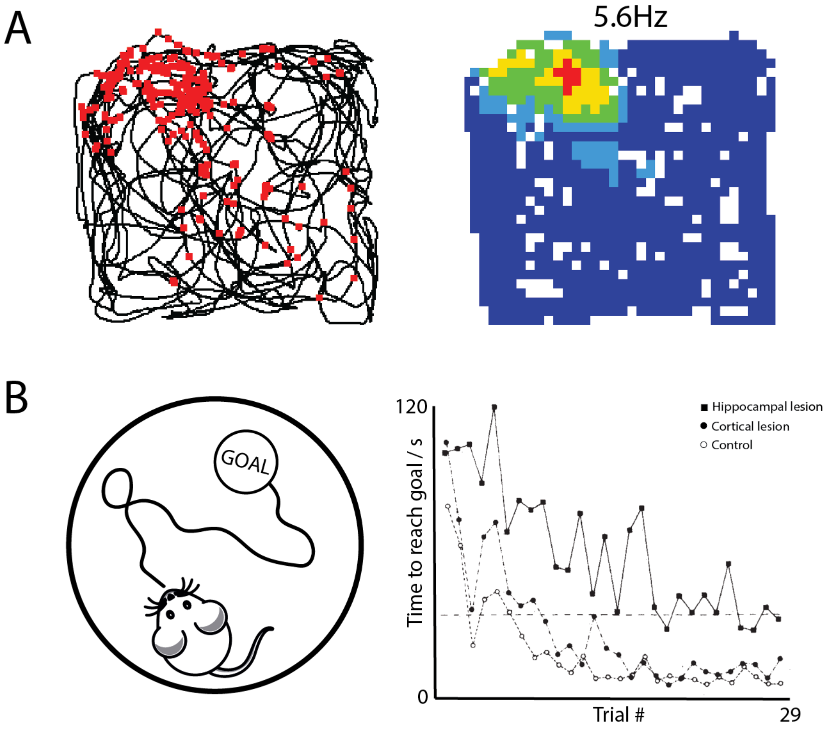


Figure 1.2: Preliminary evidence for a hippocampal cognitive map. (A) An example place cell recorded from the hippocampus. The black line indicates the trajectory of the animal and the red dots indicate action potentials (left). By observing the occupancy-normalised, firing rate map for this cell (right), it reveals a spatially dependent receptive field (place field) that predominantly fires when the animal is in top left corner of the environment. Figure adapted from Barry (2007). (B) Rats were made to navigate to a goal location in a circular open field (left). Animals that had their hippocampus legioned were significantly worse at this task than animals with cortical legions or healthy controls (right). Figure adapted from Morris et al. (1982).

This behavioural evidence for a representation of space in the brain led researchers to investigate as to how and where it might be stored. It was when O'Keefe and Dostrovsky (1971) discovered place sensitive cells in the hippocampus of freely-moving rats (figure 1.2A) that we began to form an understanding of the neural underpinnings of space. O'Keefe and Nadel (1978) later proposed that the hypothesised cognitive map resides in the hippocampus, a theory that was later backed up by impaired place navigation in rats with hippocampal lesions (figure 1.2B; Morris et al., 1982)

Since then the hippocampus has subsequently become the most intensively studied region of the brain. Consequently, it has been shown to host a variety of spatially sensitive cells that are believed to constitute towards a neural representation of space.

1.1.2 The hippocampal representation of space

1.1.2.1 Overview

The hippocampus is a uniquely organised brain structure buried deep within the temporal lobe of the mammalian brain. Its name derives from the Latin word for seahorse, which it has historically been likened to in shape. Over the years, it has gained interest from many parties due to its unique neuroanatomy and roles in memory, epilepsy, Alzheimer's disease, neurogenesis, spatial navigation and more. Loosely speaking, when viewed along the longitudinal-axis the hippocampal formation resembles two interlocking 'C's of cell laminae - a pattern that is remarkably conserved across mammals (figure 1.3).

The hippocampal formation is comprised of 6 main regions: the hippocampus proper, dentate gyrus, subiculum, presubiculum, parasubiculum and the entorhinal cortex, with the hippocampus proper further divided into 3 main subregions: CA1, CA2 and CA3 (Andersen et al., 2006). The acronym *CA* originates from *Cornu Ammonis*, which when translated from Latin means *horn of Amun* - Amun being an Egyptian god with the head of a ram. Thus reflecting the long historic interest in the unusually curved appearance of the hippocampus.

The hippocampal formation is characterised by a unique connectivity between its regions. Unlike the reciprocal connections that have been observed between brain

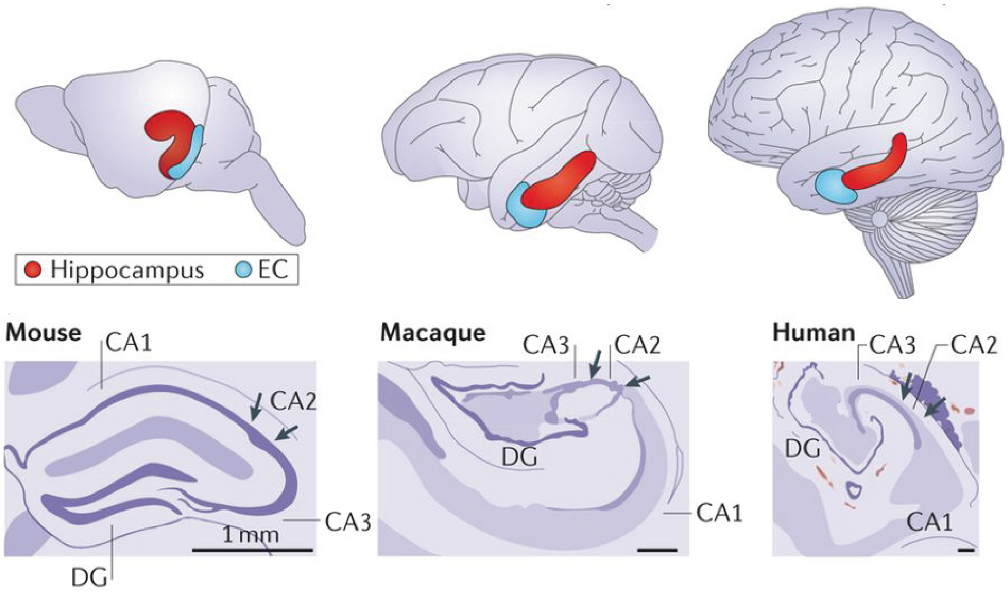


Figure 1.3: Conservation of the hippocampal formation across mammals. The hippocampal formation is buried within the temporal lobe of the mammalian brain. Columns show the location and a longitudinal cross section of the rodent, monkey and human hippocampus. Figure adapted from Strange et al. (2014)

regions in the neocortex (Felleman and Van Essen, 1991), the various hippocampal regions have a fairly distinct unidirectional connectivity between them. This was described in early drawings by Ramon y Cajal (1911) and summarised by Andersen et al. (1971) as the *trisynaptic loop* (figure 1.4). The trisynaptic loop consists of unidirectional projections from the entorhinal cortex → dentate gyrus, dentate gyrus → CA3 and CA3 → CA1. From CA1, the pattern of intrinsic connections becomes more complex, with CA1 projecting to both the entorhinal cortex and subiculum, while the subiculum also projects to the entorhinal cortex, as well as the presubiculum and parasubiculum. The return of this signal back to the entorhinal cortex thus predicates an inherent, directed recurrency in hippocampal information processing.

This unique intrinsic connectivity may be why the hippocampal formation plays such an important role in many cognitive functions - including the neural representation of space. Since the initial discovery of place cells (O'Keefe and Dostrovsky, 1971), a variety of spatially sensitive neurons have been found in the hippocampal

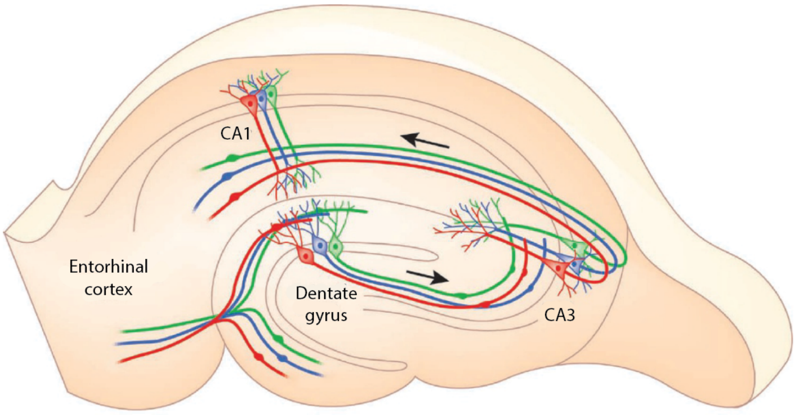


Figure 1.4: The hippocampal 'trisynaptic loop'. It is made up of unidirectional, excitatory connections that originate and terminate in the entorhinal cortex. The trisynaptic loop is comprised of projections from entorhinal cortex → dentate gyrus, dentate gyrus → CA3 and CA3 → CA1. CA1 subsequently projects back to entorhinal cortex closing the loop. Figure adapted from Moser (2011)

formation, which are believed constitute towards the neural representation of space. We will now introduce the key cell types that will be covered in this thesis.

1.1.2.2 Place cells

Place cells were discovered by O'Keefe and Dostrovsky (1971) after recording from neurons in CA1 of the rat hippocampus as it moved freely around a space. As the name suggests, place cells have a spatially tuned firing field called the *place field*, and cell activity is predominantly driven by the animal's position with respect to the place field.

When they were discovered, there were four main reasons suggesting that the firing of place cells represented a more abstract concept of place rather than simple sensory stimuli (O'Keefe, 1976). First, it does not seem to be possible to isolate any single sensory stimuli that reliably controls a cell's place field. Second, if the rat has some experience of exploring the environment in the light, then the place fields of place cells are generally the same in darkness with the lights turned off (Quirk et al., 1990). Third, the activity of place cells in open environments does not seem to be affected by the heading direction of the animal (Muller et al., 1994). Finally, place cell firing does not seem to be affected by the motivation of an animal - whether it be in search of food, water or exploring an unrewarded environment. While exceptions to these 'rules' of place cell firing have been discovered, generally they can be considered accurate.

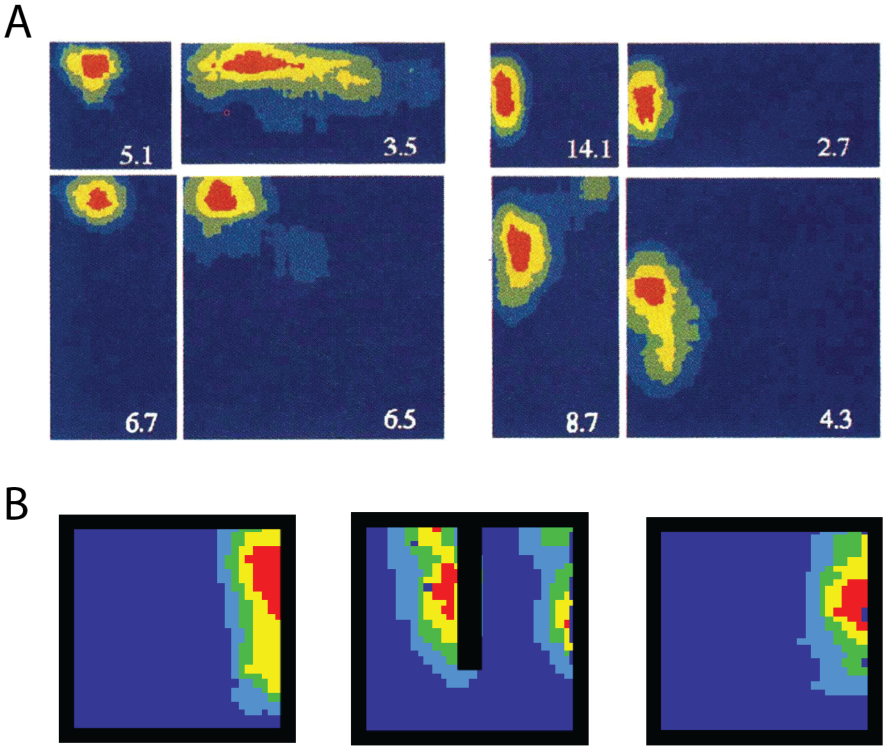


Figure 1.5: Place cells are driven by environmental boundaries. (A) O'Keefe and Burgess (1996) recorded from CA1 place cells in an experiment where they elongated or compressed one or both dimensions of a rectangular enclosure. They found that the place fields appear to be modulated by this environmental distortion in a manner that suggests that the distance and direction to boundaries has a strong influence on place cell firing. Figure adapted from O'Keefe and Burgess (1996). (B) This boundary-related influence is further demonstrated by insertion of a barrier into an environment which often causes place fields to duplicate. Duplicated fields subsequently disappear when the boundary is removed. Figure adapted from Lever et al. (2002).

The place fields of place cells do not transcend physical boundaries and their size and shape tends to vary with the size and shape of the animal's enclosing space (Muller and Kubie, 1987). This was investigated more in an experiment by O'Keefe and Burgess (1996) where they systematically elongated or compressed one or both dimensions of an open field environment (figure 1.5A). They found that place fields generally respond to the environmental distortion by elongating or compressing in a commensurate fashion. Furthermore, the location of a place field before and after the change appeared to be largely driven by the distance and direction to environmental boundaries.

The effect of environmental boundaries on place cell activity was further studied by Lever et al. (2002) in an experiment where they examined the place cell response to the insertion of a partially dividing barrier to an open field environment (figure 1.5B). They found that the barrier insertion caused numerous place fields to duplicate either side of the dividing barrier. Importantly, removal of this barrier caused the place fields to revert back their previous state before exposure to the barrier.

1.1.2.3 Grid cells

While place cells appear to form a map of positions in an environment, it is not clear how this representation would be capable of calculating vector-based optimal shortcuts through unexplored space like Tolman (1948) observed in rats (figure 1.1B). This led researchers to believe that there should be cognitive representation of Euclidean space (O'Keefe and Nadel, 1978).

Grid cells in the entorhinal cortex were discovered by Hafting et al. (2005) and are believed to somewhat facilitate this Euclidean representation of space. As with place cells, grid cells have spatially sensitive firing fields that respond to the animal being in particular location. However a given grid cell will have multiple firing fields, called *grid fields*, and they are arranged in a tessellating hexagonal lattice (figure 1.6A). The striking regularity of these grid fields is stable within environments and the relative spacing between adjacent grid fields is maintained when an animal is moved from one familiar environment to another (Hafting et al., 2005). Furthermore, the arrangement of grid fields in an environment does not seem to be affected by the animal being in darkness with the lights turned off.

Due to the robust regularity of their firing fields, grid cells are usually characterised by three metrics: the *grid orientation*, *grid scale* and *grid phase* (figure 1.6B). The

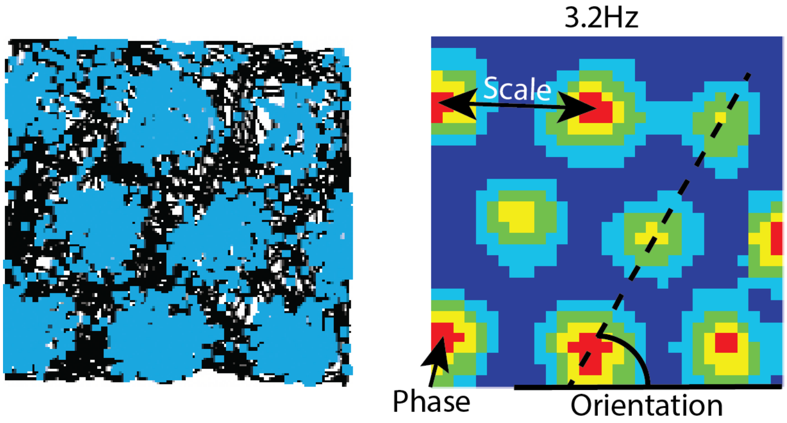


Figure 1.6: Properties of grid cells. (A) As the animal moves through space (black line), the distribution of grid cell action potentials (blue dots) forms a spatially periodic hexagonal lattice. (B) This becomes even more evident when converted to a occupancy-normalised firing rate map. The hexagonal grid can be described by its orientation, scale and phase as indicated. Figure adapted from Barry (2007).

grid orientation of a cell is defined as the angle between the alignment of grid fields and an arbitrary axis. The *grid scale* is the distance between adjacent grid fields, and therefore defines the spatial frequency of the regular pattern. Finally, the *grid phase* is the position of the regular pattern in the x-y plane, and thus can be used to calculate the relative offset between a pair of grid cells. Nearby grid cells in the entorhinal cortex have been found to have the same orientation and scale, whilst tiling the entire space via various phase offsets (Barry et al., 2007; Stensola et al., 2012). Meanwhile, the scale of the grid patterns has been shown to increase down the dorsal-ventral axis (Brun et al., 2008). In addition to these neuroanatomical findings, a hexadirectional modulation of BOLD signal has been observed in the human entorhinal cortex as participants navigate a virtual reality in an fMRI scanner (Doeller et al., 2010). This suggests a population of grid cells with similar properties to those observed in rodents may exist in the human brain.

In order measure the hexagonal regularity of a grid cell's firing fields, Sargolini et al. (2006) introduced a measure called *gridness*. The gridness of grid cell is a measure of it's 6-fold symmetry. It is calculated using the spatial autocorrelogram r of grid cell's rate map λ, which can be computed by taking the Pearson's correlation between the rate map and itself at a spatial offset (τ_x, τ_y):

$$r(\tau_x,\tau_y) = \frac{n\sum_{x,y}\lambda(x,y)\lambda(x-\tau_x,y-\tau_y) - \sum_{x,y}\lambda(\tau_x,\tau_y)\sum\lambda(x-\tau_x,y-\tau_y)}{\sqrt{n\sum_{x,y}\lambda(x,y)^2 - (\sum\lambda(x,y))^2}\sqrt{n\sum_{x,y}\lambda(x-\tau_x,y-\tau_y)^2 - (\sum\lambda(x-\tau_x,y-\tau_y))^2}} \tag{1.1}$$

where n is the number of overlapping bins in the two offset copies of the rate map. After isolating the annulus containing the 6 peaks surrounding the central peak, the Pearson's correlation is calculated between this annulus and itself rotated at a range of angles: $[30°, 60°, 90°, 120°, 150°]$. The gridness is defined as the difference between the maximum correlation coefficients for $[60°, 120°]$ and the minimum correlation coefficient for $[30°, 90°, 150°]$.

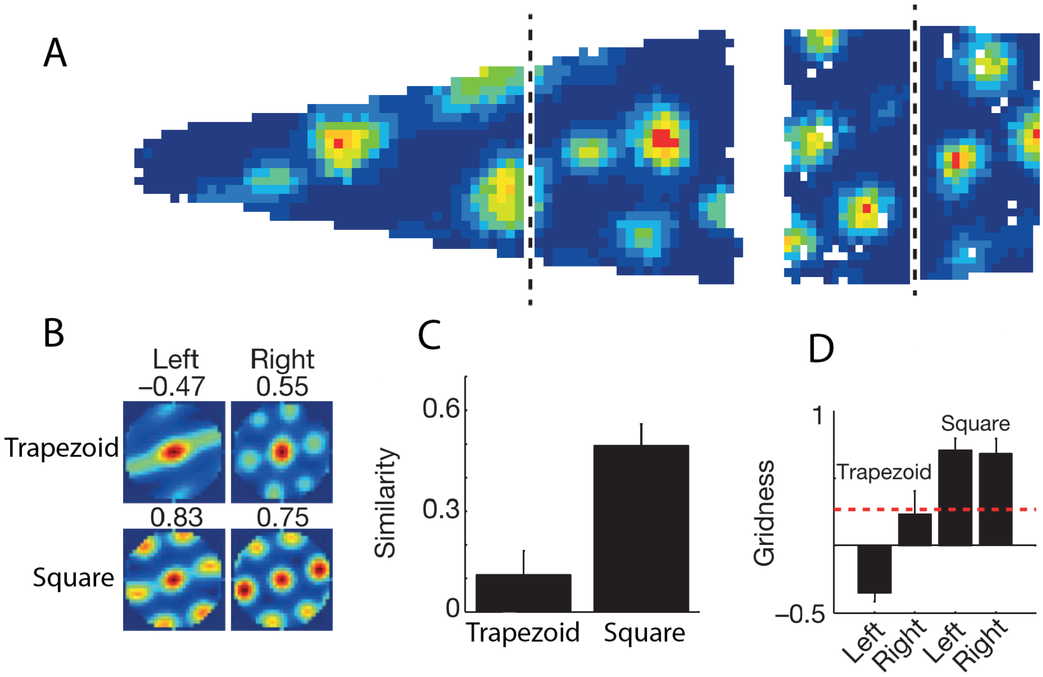


Figure 1.7: Environmental geometry influences the grid cell firing patterns. (A) Krupic et al. (2015) recorded from grid cells in the entorhinal cortex of rats as they explored a square and a trapezoidal enclosure. These firing rate maps were divided into two halves (indicated by the white line) and the spatial autocorrelograms were computed for each half (B). (C) Not only were the autocorrelograms more similar between the two halves of the square environment, but the gridness of the grid cell firing patterns recorded in the narrow end of trapezoid were significantly lower (D). Figure adapted from Krupic et al. (2015).

Whilst the scale of a grid cell appears to be relatively conserved in environments of different sizes (Hafting et al., 2005), if the environment is stretched in a manner similar to O'Keefe and Burgess (1996) the grid cells show a corresponding stretch in scale along the stretched dimension (Barry et al., 2007). Furthermore, the repetitive firing patterns of grid cells have been shown to be heavily influenced by the shape of the environment. Krupic et al. (2015) recorded grid cells in a variety of shaped enclosures and found that grid cells recorded in highly polarised environments such as trapezoids were significantly less regular than the same grid cells recorded in square environments (figue 1.7). On top of this, the grid field patterns in the narrow end of the trapezoid had significantly lower gridness than in the broad end. Behavioural correlates of these grid distortions have been observed in the spatial memory of human participants learning object locations in square and trapezoidal environments (Bellmund et al., 2019).

Due to the striking hexagonal symmetry of grid cell firing fields, one might wonder *how* and *why* they exist naturally in the brain. As for *how*, it has been shown that by using dimensionality reduction techniques on an ensemble of synthetic place cells, one is able to extract a set of basis features that look remarkably similar to the firing patterns observed in grid cells (Dordek et al., 2016; Stachenfeld et al., 2017). Furthermore, as to *why* these patterns exist, it has long been hypothesised that grid cells might be important for solving path integration problems in spatial navigation - that is *where am I with respect to a previous location?* Not only would a low dimensional representation of place cells be an ideal basis for updating a place cell map using self-motion cues, but recent work studying mice navigating in a 2d virtual reality has been able to pull apart the sensory and path integration components of spatial navigation. Consequently, it has been shown that following a mismatch between the virtual scene and motor input, grid cell populations show a more consistent mapping to the motor coordinate interpretation of space, whereas place cells show a more reliable map when viewed with respect to the visual scene (Chen et al., 2019).

1.1.2.4 Boundary vector cells

The observation that CA1 place cells were strongly driven by environmental boundaries led researchers to hypothesise that there should be cells in nearby areas of the hippocampal formation that were explicitly driven by boundary responses (O'Keefe and Burgess, 1996; Hartley et al., 2000; Barry et al., 2006).

Boundary vector cells were discovered in the subiculum of the hippocampal formation by Lever et al. (2009). As their name suggests, they are defined by a receptive field with maximal firing when a boundary is at a particular distance and direction from the rat (figure 1.8). Similar cells with shorter distance tunings have been found in the entorhinal cortex (Solstad et al., 2008) as well as the presubiculum and parasubiculum (Boccara et al., 2010).

Generally boundary vector cells can be thought of as the conjunction of two tunings; a directional tuning with a preferred allocentric direction from the animal to the boundary, and a distal tuning with a preferred distance from the animal to the boundary. An important note is that boundary vector cells respond to both internal and external boundaries in the environment. Furthermore, they respond predictably to the insertion of a new boundary into the environment and in the darkness with the lights turned off (Lever et al., 2009).

A sensible question would be *what constitutes a boundary?* Evidence suggests that what the animal perceives as a boundary is an abstract concept dependent on both appearance, texture and impediment to movement. They have been shown to respond regardless of the colour, material or shape of the boundary (Lever et al., 2009). Furthermore, they also respond to vertical drops in an environment such as the separation between two platforms - even when the gap is small enough that the other platform is reachable.

Boundary vector cells are quite unique in the sense that they were predicted from the properties of place fields over a decade before they were eventually discovered.

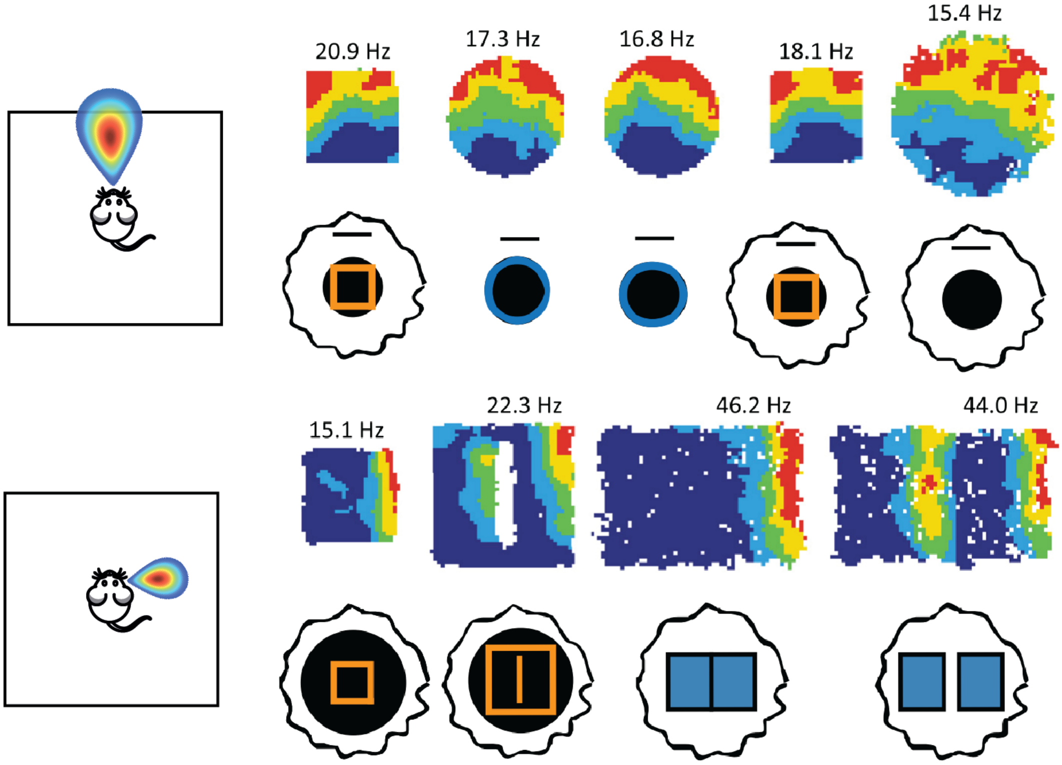


Figure 1.8: Boundary vector cells recorded from the subiculum of the hippocampal formation. Leftmost of each row shows the boundary vector cell receptive field which fires maximally at a preferred distance and allocentric direction to a boundary. The rest of the row shows shows the firing rate map for the corresponding boundary vector cell in a variety of different environments. Figure adapted from Lever et al. (2009).

O'Keefe and Burgess (1996) first proposed the idea that place fields could be modelled by a sum of boundary driven responses, before Hartley et al. (2000) formalised the proposal into the boundary vector cell model of place cell firing.

1.1.3 The boundary vector cell model of place cell firing

The boundary vector cell model of place cell firing proposes that place fields are a function of the environmental boundaries relative to the animal. Specifically, it proposes that the firing of a place cell in a given location is driven by feedforward connections from a set of boundary vector cell inputs. The receptive field of a boundary vector cell is defined as the product of two gaussians; one tuned to a specific distance and another tuned to a specific allocentric direction (figure 1.9A).

More specifically, the firing for a boundary vector cell (with preferred distance d and angle ϕ) to a boundary at distance r and direction θ, subtending at an angle $\delta\theta$ to the animal is given by:

$$\delta f(\boldsymbol{x}) = g(r,\theta)\delta\theta \tag{1.2}$$

where $\boldsymbol{x}$ is the animals location and:

$$g(r,\theta) \propto \frac{\exp[-(r-d)^2/2\sigma_{rad}(d))]}{\sqrt{2\pi\sigma^2_{rad}(d)}} \times \frac{\exp[-(\theta-\phi)^2/2\sigma_{ang}]}{\sqrt{2\pi\sigma^2_{ang}}} \tag{1.3}$$

In the model, the angular tuning width σ_{ang} is constant and radial tuning width increases linearly with the preferred tuning distance: $\sigma_{rad}(d_i) = d_i/\beta + \xi$ for constants β and ξ.

It proposes that the firing F of a place cell is proportional to the thresholded sum of its n boundary vector cell inputs (figure 1.9B).

$$F(\boldsymbol{x}) \propto \Theta(\sum_{i=1}^{n} f_i(\boldsymbol{x}) - T) \tag{1.4}$$

where T is the place cell's threshold and

$$\Theta(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1.5}$$

As with real boundary vector cells, the directional component of their firing is calculated in an allocentric reference frame and is therefore independent of the animals heading direction. Consequently, modelled boundary vector cell firing is only de-

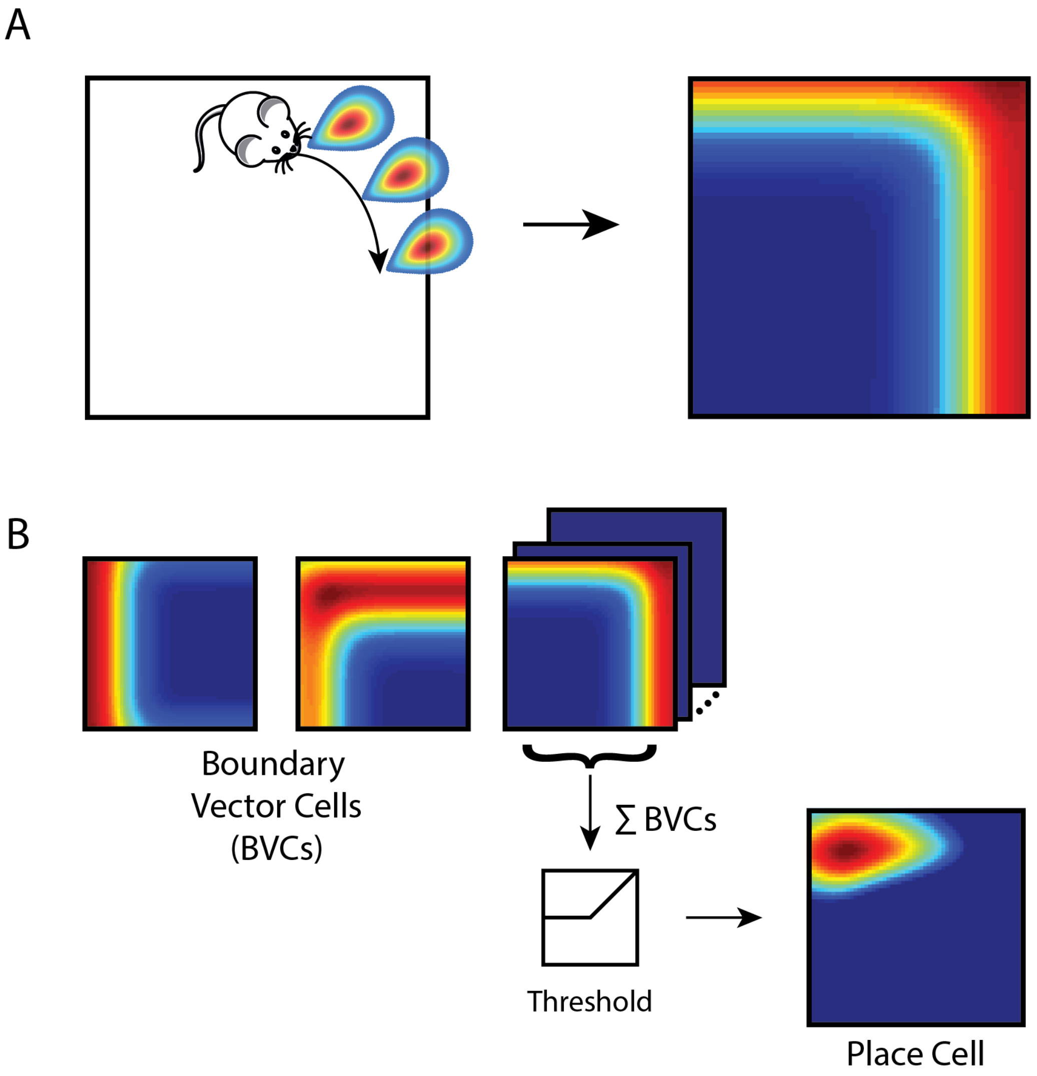


Figure 1.9: The boundary vector cell model of place cell firing. (A) Boundary vector cells are characterised by a tuning to a preferred distance and allocentric direction to boundaries. Consequently, the firing of boundary vector cells is independent of the heading direction of the animal. (B) Place cell firing in a particular location is modelled as the thresholded sum of its feedforward boundary vector cell inputs.

pendent on the environmental boundaries and the animals position.

The model fits the observed distortions of place fields in the experiment by O'Keefe and Burgess (1996) (figure 1.5A) where one or both dimensions of the environment were elongated or compressed. The framework also predicts the spatial memory of human participants remembering the location of objects in a virtual environment that was subjected to the same elongations and compressions (Hartley et al., 2004).

Boundary vector cell models have since been used to explain place cell responses to barrier insertions (Barry et al., 2006; Barry and Burgess, 2007) and multicompartmental environments (Grieves et al., 2018).

1.1.4 Summary

In this section we introduced the spatial navigation problem and the neural representation of space in the hippocampal formation. Despite describing this representation, we did not go into any great detail of how it might be used to guide action selection in order to solve spatial navigation problems. Indeed, the process of choosing appropriate actions in order to reach goals strays into the realm of machine learning theory, and brings us to reinforcement learning.

1.2 Reinforcement learning

1.2.1 What is reinforcement learning?

Reinforcement learning is about learning what actions to take in an environment so as to maximise a numerical reward signal. This reward signal can take any real value, with positive values typically indicating rewarding experiences and negative values typically indicating unrewarding experiences. The magnitude of this signal at a point in time therefore defines how rewarding or unrewarding the current experience is.

Unlike types of supervised learning, the learner is not told what actions it should take. Rather it is left to choose by monitoring the reward signal during some form of trial-and-error search. In particular, the learner will often have to negotiate the concept of delayed rewards - that some key actions may not be rewarded immediately and thus their importance must be integrated through time in order to learn effectively. Consider this example of delayed reward: you are out enjoying an evening in the pub and drinking alcohol with your friends, then you go home and sleep.

The following morning you awaken with a hangover. Without propagation through time, the negative reward of being hungover would be inappropriately associated to you waking up, rather than the amount of alcohol you drank the night before. Furthermore, the nature of the representation used to facilitate learning will have a big impact on the efficiency of the trial-and-error search. To continue with the hangover example, say you noticed a pharmacy on your way back from the pub. You may think nothing of this at the time, but in the context of being hungover you are able to exploit this seemingly unnecessary knowledge from the night before and buy some painkillers to ease the discomfort. Whilst this may seem obvious in the context of animal behaviour, this sort of preemptive consolidation of potentially relevant relationships is critical for transferring useful knowledge to unknown future tasks, and is non-trivial to implement in a computational framework.

Reinforcement learning also addresses separate issues to unsupervised learning. Whilst they may appear superficially similar due to the absence of a supervisory signal, reinforcement learning seeks to maximise the reward signal accumulated from an environment whilst unsupervised learning seeks to identify underlaying latent structure. Indeed, identifying latent structure from experience could be highly beneficial for a reinforcement learning algorithm, but it does not directly address the crux of a reinforcement learning problem which is how to choose actions in order to maximise reward. Reinforcement learning should therefore be considered a distinct type of learning alongside supervised and unsupervised learning (Sutton and Barto, 2018).

The learner and decision maker in reinforcement learning problems is called the *agent*. Everything external to the agent is referred to as the *environment*. The agent and the environment interact continually in an asynchronous loop, with the agent selecting actions and the environment responding by presenting the agent with the reward signal and a new situation (figure 1.10).

Reinforcement learning explicitly sets out to address the problem of how a goal-

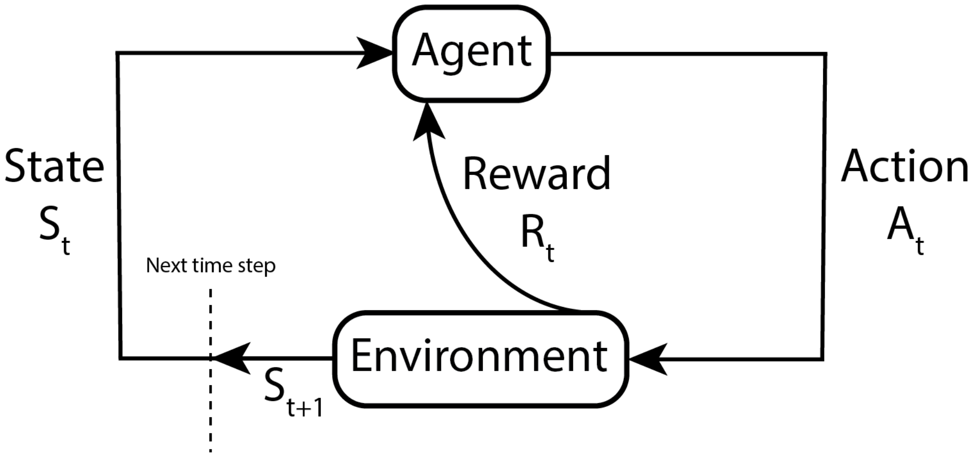


Figure 1.10: The Agent-Environment interface. The agent is presented with a current state S_t and takes a particular action A_t. The environment then responds by presenting the agent with a reward R_t and subsequent state S_{t+1} for taking that action. S_{t+1} then becomes the current state and the cycle continues.

directed agent should interact with it's environment. When this is applied to the scenario where the agent's environment consists of positions in space, and the agent's actions move it between these positions, then solving the reinforcement learning problem becomes equivalent to solving a spatial navigation problem.

1.2.2 Elements of reinforcement learning

The specification of a reinforcement learning problem requires three components: a *state space*, an *action space* and the *reward signal.*

The *state space* is a set of variables used to mathematically summarise the agent's state of being in the environment. In the simplest scenarios such as discrete cases, these states are binary and non-overlapping. For example, in a spatial setting one could use the grid lines on a map to form a discretised representation of the environment - each square formed by the grid would correspond to a state in the state space (figure 1.11). The agent's current state of being in the environment can therefore be summarised by the set of binary variables indicating whether or not the agent is in each of these squares. Alternatively the state space could be continuous; for example the latitude and longitude coordinates on the map, or the firing rates of spatially sensitive neurons.

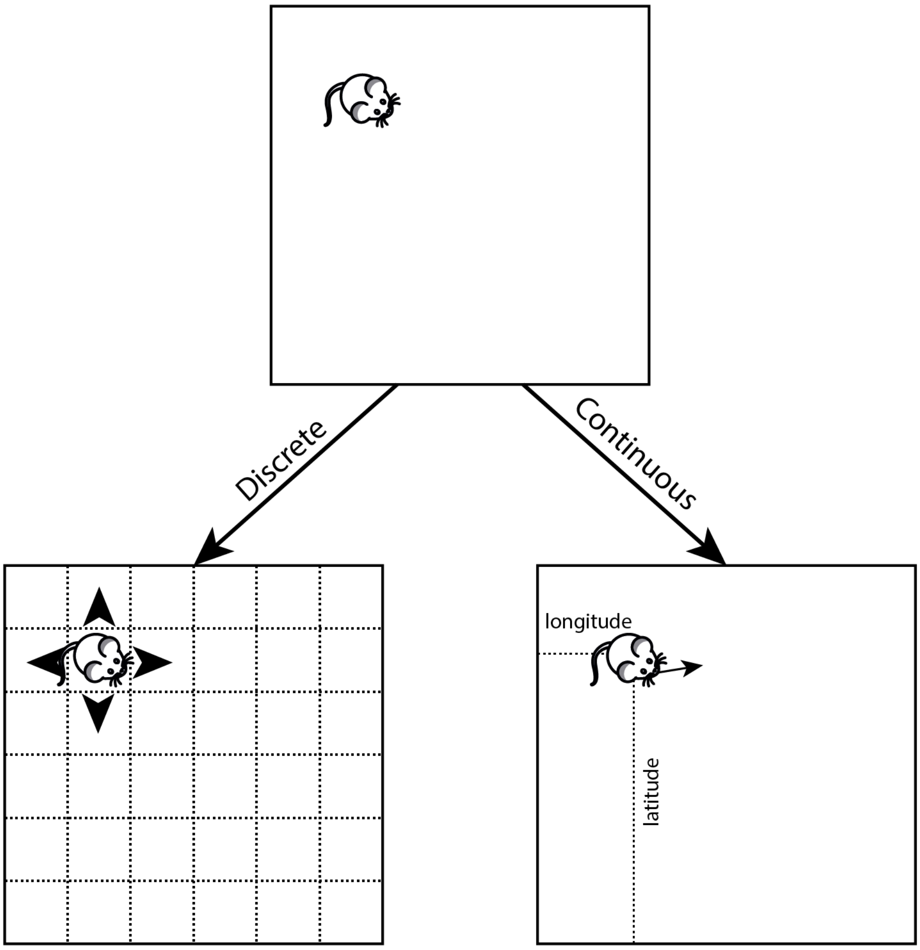


Figure 1.11: Discrete and continuous state and action spaces for a spatial navigation reinforcement learning problem. The position of the animal in space (top) can be implemented in a reinforcement learning framework with a discrete (left) or continuous (right) state action space. In the discrete scenario, positions in space are described by a non-overlapping grid of squares - the states. The actions available to the agent (indicated by the arrows) allow it to move between adjacent states. In the continuous scenario position is summarised by a set of coordinates describing latitude and longitude location. Similarly, the action-space consists of a vector describing direction of movement in the coordinate reference system.

The *action space* is a set of variables used to transition through the state space. In the discrete case, the agent chooses which distinct action to use from a finite set - for example whether to move to the North, South, East or West adjacent states on the map. In a continuous action space, actions are summarised by a single, real-valued vector - for example the direction in which to move on the map expressed by its longitudinal and latitudinal components, with a magnitude proportional to the agent's speed/acceleration.

The *reward signal* defines the goal of the reinforcement learning problem. At each time step, the environment sends the agent a single real-valued number called the *reward*. The aim of the agent is to maximise the total reward it receives from the environment over time by its choice of actions in the state space. Consequently this reward signal defines what the agent perceives as good or bad choices. In a biological system, this reward signal can be thought as analogous to the experience of hedonistic pleasure (e.g. good choice) or suffering (e.g. bad choice).

Whilst the state space, action space and reward signal defines a reinforcement learning problem, specifying a solution to the problem requires two more components: a *value function* and a *policy*.

Though the reward signal provides an immediate indication of whether something is good or bad, the *value function* is a way of expressing what is good or bad in the long run. More precisely, the value V of a particular state s is the amount of reward r that the agent can expect to accumulate starting from state s, with a temporal discounting of distant rewards according to a discount parameter $\gamma \in [0,1]$.

$$V(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s] \tag{1.6}$$

A discount parameter $\gamma = 1$ indicates a maximally long-sighted value function that favours long-term rewards just as much as short-term rewards, whereas a $\gamma = 0$ describes a maximally short-sighted value function that only cares about the next immediate reward. Therefore for $\gamma > 0$, it is possible for a state to have a low reward but a high value if it usually precedes highly rewarding states. The opposite is also true where an initial state might be quite rewarding (e.g. drinking alcohol) but can have low value if it usually leads to highly unrewarding states (e.g. hangover). Thus, whereas the reward determines the immediate desirability of environmental states, the value determines the long-term desirability of those states by taking into account the temporal relationship between states.

The *policy* defines the way of choosing actions in a particular state at a given time, and therefore describes a mapping between the state and the action space. Since the aim of reinforcement learning is to act in a way that optimises expected future reward, the policy is usually defined in terms of the value function for potential future states, or the actions leading to them. For example, one could define a greedy policy that always chooses the action that leads to the next available state with the highest value. Alternatively, an agent might want to sample their next state probabilistically in a manner that is proportional to their values. In fact, it is important for policies to be somewhat stochastic in order to avoid reaching a local minima; what is known as the exploration-exploitation tradeoff. The agent must exploit the knowledge that it has attained from prior experience in order maximise future reward, nevertheless it must also explore in case there are more rewarding states in the environment that it has yet to experience, or if the reward contingencies have changed.

There is an optional, third component that can be used to specify a solution to a reinforcement learning problem which is an agent's model of the environment. The intention behind such a model is that it somewhat imitates the structure of the environment in a way that can facilitate planning. By planning we refer to a procedure that considers possible evolutions of the agent's future state in order to help guide actions. Therefore, if the policy is viewed as a mapping between states and actions, a model can serve as an intermediary step where states are mapped onto the model which subsequently informs action selection. Consequently, solutions that use such models are known as *model-based* algorithms. These are in contrast to so-called *model-free* algorithms that do not use a model, and thus rely more directly on trial-and-error experiences to gauge optimal actions. The extent to which an algorithm relies on an internal model of the environment or the caching of trial-and-error experience can be viewed as a spectrum of possible reinforcement learning solutions. These range from the high-level, deliberative planning of model-based methods to the low-level, value-caching methods of model-free solutions.

1.2.3 Types of reinforcement learning

The extent to which a reinforcement learning algorithm relies on a model affects how it can react to changes in the transition and reward structure of the environment. For example, after learning that reward in the environment has moved from one location to another, a model-based system would be able to reuse the model of the environment and alter the planning procedure to plan a route to the newly rewarded location. On the other hand, a model-free system needs to learn from trial-and-error that the reward has moved locations, until it eventually devalues the old goal location enough to permit actions to the new one. For this reason, model-free strategies are considerably less flexible to adapt to changes in the environment. The upshot of this is that they are considerably more efficient than their model-based counterparts due to the computational overhead imposed by a planning procedure. Importantly, both model-free and model-based methods will converge upon the same set of optimal policies eventually, and so it is only through perturbing the reward or transition structure of the environment that an observer could pull their behaviour apart. Such perturbations to the reward structure in the environment is known as *reward revaluation*, whereas perturbations to the transition structure is called *transition revaluation*.

We will now go through these types of reinforcement learning in greater detail, and then introduce an algorithm that somewhat spans the void between a model-free and model-based approach called the successor representation.

1.2.3.1 Model-free

Since a hallmark of model-free approaches is a heavy reliance on trial-and-error, it is important for them to be able to keep track of the respective outcomes from these experiences. Considering that the ultimate goal of a reinforcement learning agent is to maximise reward accumulation, a sensible quantity for a model-free agent to keep track of is the value of states (equation 1.6). Crucially, it is possible to iteratively decompose the value of a current state $V(s_t)$ in terms the expected reward r_t and value associated with the next state. This is known as the Bellman equation:

$$\begin{aligned} V(s_t) &= \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + ...] \\ &= \mathbb{E}[r_t + \gamma V(s_{t+1})] \end{aligned} \tag{1.7}$$

This means that in an unchanging environment it possible to continually improve a value estimate of a state using a temporal-difference learning rule, which utilises the difference between predicted outcomes and the actual outcomes to improve the accuracy of the predicted estimate (Sutton, 1988).

$$V(s_t) \leftarrow V(s_t) + \alpha[r_t + \gamma V(s_{t+1}) - V(s_t)] \tag{1.8}$$

Where $\alpha \in [0,1]$ is a learning rate parameter that dictates how much we adjust our initial estimate $V(s_t)$ in light of a new estimate $r_t + V(s_{t+1})$ made using the Bellman equation (1.7).

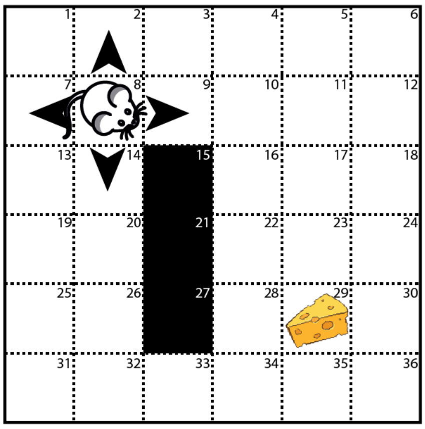
Furthermore, since knowing all $V(s)$ would still require knowledge of how the states are connected in order to choose the appropriate action, an even more model-free approach would be to instead estimate the value of taking a particular action a in a particular state s. These so-called *action-value* functions are denoted by $Q(s,a)$ (figure 1.12) and are related to the state-value function according to the relationship:

$$V(s) = \sum_a \pi(a|s) Q(s,a) \tag{1.9}$$

Here $\pi(a|s)$ is the probability of the agent taking action a given it is in state s and thus represents the agent's *policy*. In order to formalise an update equation for Q that is analogous to equation (1.8), one must first choose this policy. A popular model-free algorithm that has been shown to converge to the optimum action-value function is Q-learning (Watkins and Dayan, 1992), which uses a greedy policy to estimate the value of the next state $V(s_{t+1}) \approx \max_a Q(s_{t+1}, a)$:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \qquad (1.10)$$

To negotiate the exploration-exploitation dilemma, this is often paired with an ε-greedy policy to decide actions - where the agent chooses a random action with probability $\varepsilon \in [0, 1]$ and chooses the greedy action $a = \text{argmax}_a Q(s, a)$ otherwise.



$$Q = \text{States} \begin{array}{c|c|c|c|c} & \multicolumn{4}{c}{\text{Actions}} \\ & N & S & E & W \\ 1 & 0 & 0.1 & 0.1 & 0 \\ \hline 2 & 0 & 0.2 & 0.2 & 0.1 \\ \hline 3 & 0 & 0.4 & 0.4 & 0.2 \\ \hline 4 & 0 & 0.7 & 0.7 & 0.4 \\ \hline & \vdots & \vdots & \vdots & \vdots \end{array}$$

Figure 1.12: Example of a model-free reinforcement learning agent. The agent updates the action-value matrix Q by monitoring the reward signal during trial-and-error search. It then uses Q to pick actions that have the most value.

Using the action-value function Q makes Q-learning completely model-free in the sense that it does not require any knowledge of the structure of the environment to work, allowing it to work well on a vast range of independent tasks (Mnih et al., 2015). However the lack of internal model means that it struggles to generalise knowledge across different goals within the state-space - old goals need to be unlearnt in order to learn a new one.

1.2.3.2 Model-based

The *model* in model-based methods constitutes to anything that the agent can use to help it predict how the environment will respond to its actions. In order to capture the randomness of the real world this model can be stochastic, in which case there are multiple ways the environment could respond, each with some probability of happening. Stochastic models can be separated into two kinds, models that unfold all possible evolutions of the environment and their respective probabilities of happening - called *distribution models*, and models that draw samples from that distribution - called *sample models*. Sample models work by drawing either a single sample, or multiple samples and combining the outcomes using a method such as particle filtering (Del Moral, 1997).

Given an initial state and policy, a sample model can simulate entire sequences of states, actions and rewards, whereas a distribution model can simulate all possible sequences and their probabilities of happening. Consequently, a distribution model can be used to generate samples, making them more powerful but harder to obtain. Either way, the model can be viewed as being used to simulate the outcome of a prospective action, or a sequence of prospective actions. This process has strong similarities to the concept of cognitive planning in psychology (Morris and Ward, 2004). In reinforcement learning, the term *planning* is used to refer to any computational process that takes a model as input and produces or improves a policy for interacting with the modelled environment (Sutton and Barto, 2018).

An important element of this definition of planning is that it can only produce or improve the policy with respect to the modelled environment. If the model is a poor fit to the environment, then the policy outputted by the planning procedure is unlikely to be of actual benefit in reality. This dependence on the goodness of the model can make model-based methods less universally applicable than model-free methods, with models often needing elements of the task structure to be hardcoded into them in order to work best, such as the chess algorithm Deep Blue (Hsu, 1999).

To apply a reinforcement learning framework in the context of spatial navigation, a sensible internal model to use use would be some form of map of the environment. For example, imagine a grid world where the aim of the agent is to navigate from a starting location to a goal in order to receive some reward (figure 1.13). The most useful internal model for the agent to have would be a replica of the actual environment. That way the policy outputted by a model-based planning procedure will be directly applicable to the real world. However, presuming the exact state of the environment is only partially observable to the agent, we need some way of updating the model based on those partial observations. For example, if the agent observes that some parts of the environment are blocked and thus those states are inaccessible from all other states, it needs a way of representing that in its model to best match the state of the environment. We can do this by using a stochastic model of the environment Ξ and implementing a learning rule to update the probability $\Xi(s')$ that a nearby state s' is accessible after observing it.

$$\Xi(s') \leftarrow \Xi(s') + \alpha[\mathbb{1}_{\Xi}(s') - \Xi(s')] \tag{1.11}$$

where α is a learning rate and the function $\mathbb{1}_{\Xi}(s')$ indicates whether the observed state s' was accessible or not:

$$\mathbb{1}_{\Xi}(s') = \begin{cases} 1 & \text{if } s' \text{ is accessible} \\ 0 & \text{otherwise} \end{cases} \tag{1.12}$$

By reflecting these partial observations in the stochastic representation of the environment Ξ, the agent could use it to randomly draw a map from this distribution of possible maps (i.e. a sample model). Then on this sampled map it could plan the shortest route to the goal. Alternatively, a distribution model would entail calculating all the possible maps and their probability of existing - an approach that would quickly become combinatorially intractable, even for moderately sized state spaces.

For this reason, we will not be covering distribution models in this thesis.

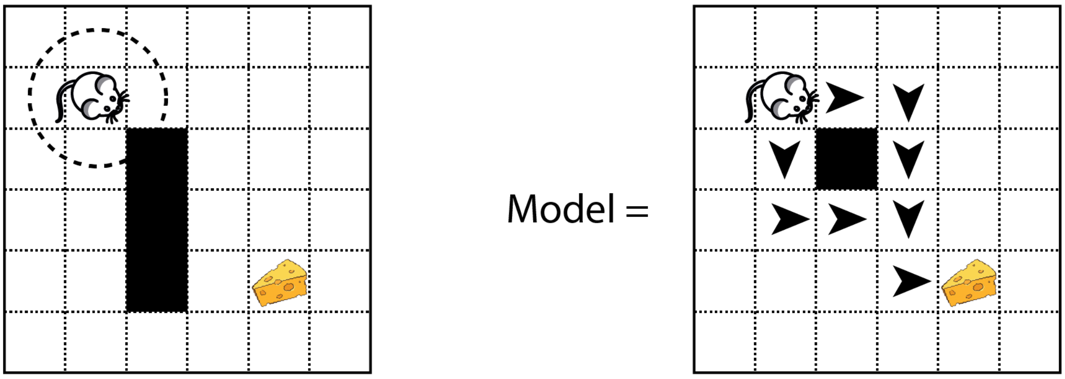


Figure 1.13: Example of a model-based reinforcement learning agent. The agent is endowed with a radius of vision (dashed circle) with which it can partially observe the environment (left). With this it updates it's model of the environment and uses it to plan a route to the goal (right).

A map-like representation such as the one just described requires privileged knowledge of the task; such as the initial map, state connectivity and that some states will become inaccessible. One of the simplest yet general ways to represent a useful model of the environment is through the *state transition probabilities*. Given you are in a particular state s and take a particular action a, the state transition probabilities define a probability distribution over next possible states s': $Pr(s_{t+1} = s'|s_t = s, a_t = a)$. In a spatial framework, the sparsity of this representation can be reduced by acknowledging the one-to-one mapping between state-action pairs and the subsequent next state $(s_t, a_t) \rightarrow s_{t+1}$. For example, if the agent is in a particular location s and takes action a to head North, there is only one state that the agent can end up in - the state to the North of s. This means we can further condense these state transition probabilities to a mapping between states and future states in the next time step. This can be represented in the *one-step transition matrix* $T(s,s') = P(s_{t+1} = s'|s_t = s)$. Rows of T therefore represent current states and columns of T represent the states that immediately succeed them, such that $T(i, j)$ is equal to the conditional probability that the agent will transition to state j in the next time step, given that it is currently at state i. Since the rows of T represent conditional probability distributions, they sum to unity: $\sum_j T(i, j) = 1$.

Using the one-step transition matrix T, one could use it to implement a sample model-based planning procedure such as a Monte-Carlo tree-search using the policy dictated by T. One could also apply T recursively to implement a distribution model looking n-time steps into the future. Alternatively, there is a way to circumvent the computational overhead associated with model-based planning by using T to directly approximate the value function. This method is called the successor representation and relies on using T to form a predictive map of the environment.

1.2.3.3 The successor representation

The successor representation (Dayan, 1993) is a way of encapsulating the short and long-term dynamics of an environment in a single matrix. It can be used to produce value estimates without the computational overhead of a planning procedure and does so by temporally discounting state occupancies as opposed to the reward signal like equation 1.6. Specifically, the successor representation matrix M can be written as the weighted sum of transition matrices:

$$\begin{aligned} M &= I + \gamma T + \gamma^2 T^2 + \gamma^3 T^3 + ... \\ &= \sum_{t=0}^{\infty} \gamma^t T^t \end{aligned} \tag{1.13}$$

Here I is the identity matrix and $T^n = \underbrace{T \times T \times ... \times T}_{n \text{ many}}$ is the n-step transition matrix, describing the transition probabilities n time steps into the future, just as T described them one time step into the future. Meanwhile, the discount parameter $\gamma \in [0,1]$ now dictates the temporal discounting of future state occupancies.

By separately learning a reward vector $\boldsymbol{R}$ such that $R(s)$ is the expected reward received upon visiting state s (figure 1.14), the successor representation is able to approximate the value function (equation 1.6).

$$V(s) = \sum_{s'} M(s,s')R(s') \tag{1.14}$$

An important feature of the successor representation is both M and $\boldsymbol{R}$ can be learnt through experience using temporal-difference learning rules. Suppose the agent moves from $s \rightarrow s'$ and receives reward r, then the agent can implement the learning rules:

$$M(s,:) \leftarrow M(s,:) + \alpha[\mathbb{1}_s + \gamma M(s',:) - M(s,:)] \tag{1.15}$$

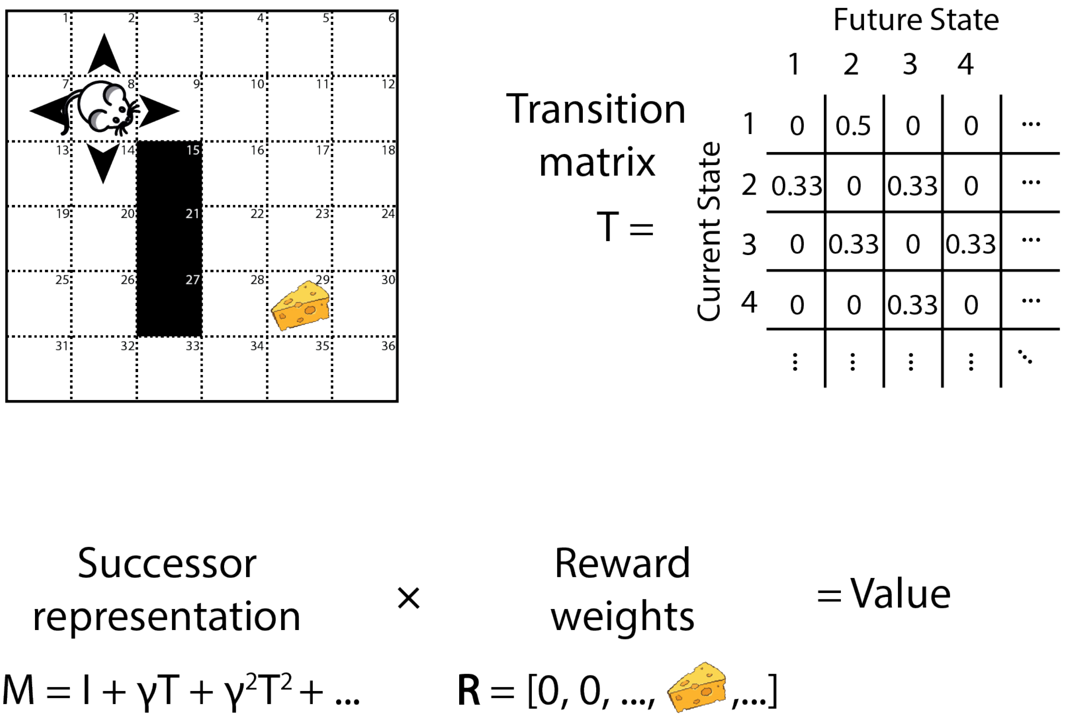
$$R(s') \leftarrow R(s') + \alpha[r - R(s')] \tag{1.16}$$

where the i$^{\text{th}}$ element of vector $\mathbb{1}_s$:

$$\mathbb{1}_s(i) = \begin{cases} 1 & \text{if } s = i \\ 0 & \text{otherwise} \end{cases} \tag{1.17}$$

and $M(i,:)$ indicates the i$^{\text{th}}$ row of the successor matrix M.

This independent learning of the successor matrix (equation 1.15) and reward weights (equation 1.16) allows the value function (equation 1.14) to be calculated by decomposing it into the transition statistics M and reward statistics $\boldsymbol{R}$ of the environment. Therefore, the successor representation can be viewed as an intermediary method between model-based and model-free. It has similarities to model-based methods since it captures the transition and reward structure of the environment, but rather than implementing a complicated planning procedure it uses these to form value estimates for action selection. Whereas it also has similarities to model-free methods in the sense that it relies on temporal difference learning rules to cache



Transition matrix

$$T = \begin{array}{c|c|c|c|c|c} & 1 & 2 & 3 & 4 & \\ \hline 1 & 0 & 0.5 & 0 & 0 & \cdots \\ \hline 2 & 0.33 & 0 & 0.33 & 0 & \cdots \\ \hline 3 & 0 & 0.33 & 0 & 0.33 & \cdots \\ \hline 4 & 0 & 0 & 0.33 & 0 & \cdots \\ \hline & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

$$M = I + \gamma T + \gamma^2 T^2 + \ldots \qquad \mathbf{R} = [0, 0, \ldots, \quad, \ldots]$$

Figure 1.14: Example of a successor representation agent. The successor representation is the weighted sum of transition matrics T, exponentially discounted into the future by a discount parameter γ. It therefore represents a temporally discounted expected future occupancy matrix. This can be multiplied the vector of expected rewards from each state $\boldsymbol{R}$ to compute value for actiopn selection.

the quantities M and $\boldsymbol{R}$ that are the precursors to value. Due to the way that the successor representation forms a probabilistic mapping of future state transitions, it is often referred to as a *predictive map*.

1.2.4 The hippocampus as a predictive map

Aside from the boundary vector cell model, another prevailing theory of hippocampal place cell firing is that the place fields encode a successor representation over potential future states in the environment (Stachenfeld et al., 2017). More specifically, it suggests that the position of the peak firing rate is the location of the state being encoded, and the firing rate in other locations is proportional to the discounted number of times that location is expected to be visited in the future. Stachenfeld et al. (2017) propose that by doing this, hippocampal place cells encode the columns of the successor matrix M. When the values in these column vectors are mapped onto their corresponding states in space, they somewhat resemble the firing fields observed in place cell rate maps.

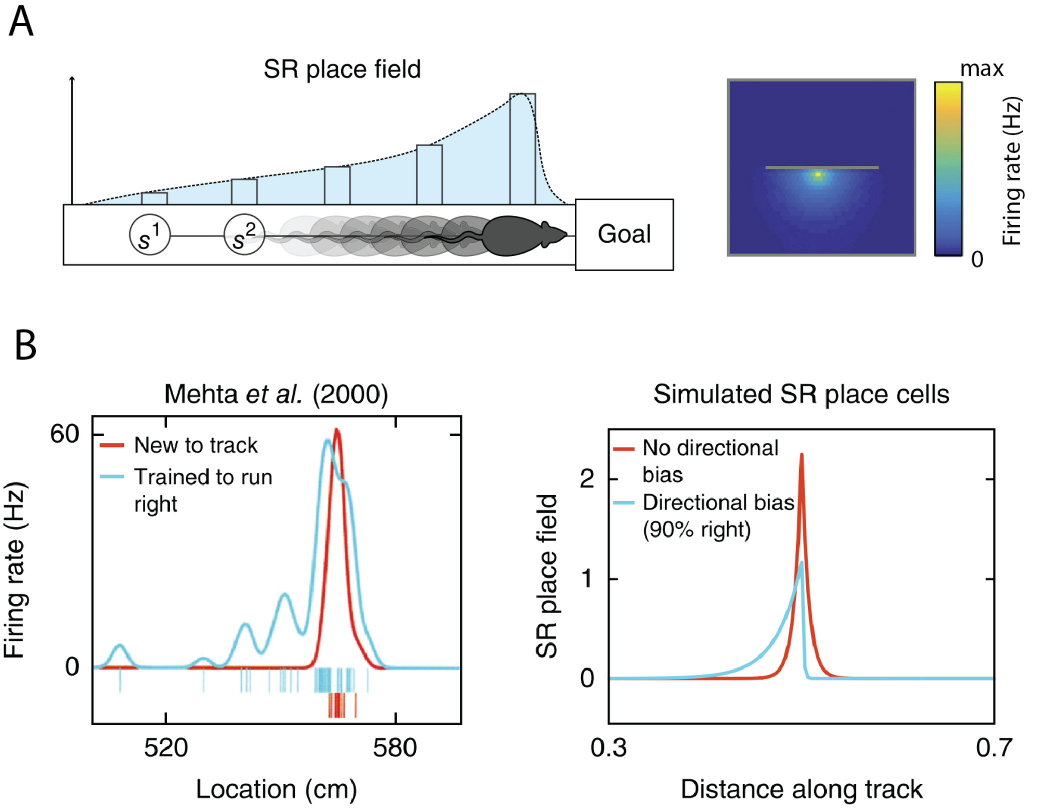


Figure 1.15: Place cells from the predictive map model of the hippocampal formation. (A) The successor representation can be used to simulate place cells in 1D and 2D environments. Like real place cells, the place fields do not extend through environmental boundaries. (B) The successor representation is able to capture the behaviour dependent skewing of hippocampal place fields on a linear track. Figure adapted from Stachenfeld et al. (2017).

Stachenfeld et al. (2017) further propose that the firing patterns of grid cells in the entorhinal cortex represent an eigendecomposition of the successor matrix M. Indeed, when the values in these eigenvectors are mapped onto the states in the environment, they exhibit spatially periodic fluctuations that bear resemblance the activity patterns of grid cells. Since M is a linear combination of transition matrices T (equation 1.13), it is useful to note that T and M share the same eigenvectors. Thus the theory proposes that entorhinal grid cells encode a basis for the transition structure of an environment. This is in accordance with recent modelling work by Dordek et al. (2016) who showed that principle component analysis (PCA) of idealised place cell activity yields principle components (PCs) that exhibit grid-like patterns. The PCA process involves calculating the covariance matrix between the simulated place cells, which resembles the transition structure of the environment via the covarying statistics of cell activities. Furthermore, the method of computing PCs often involves eigendecomposition of this covariance matrix - although Dordek et al. (2016) found that algorithms which constrain the PC's to be non-negative yielded the most hexagonal patterns.

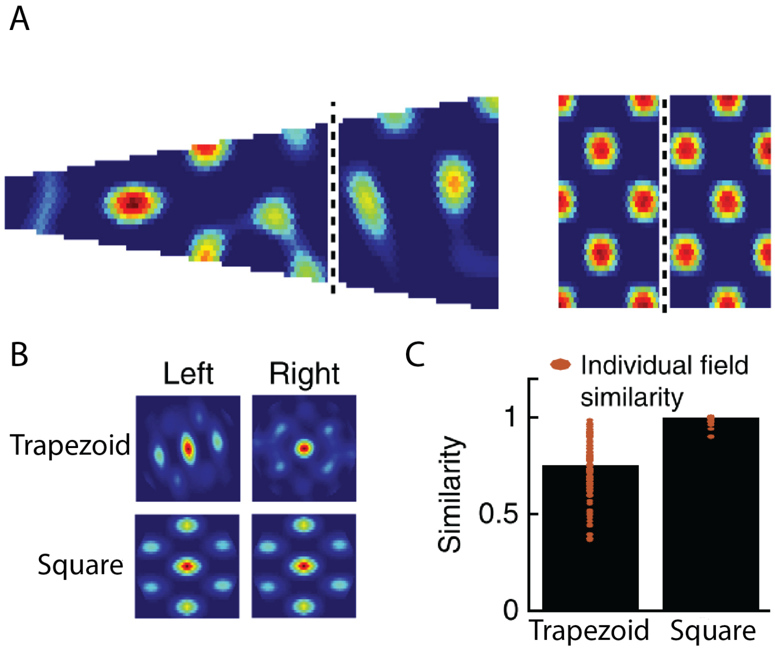


Figure 1.16: Grid cells from the predictive map model of the hippocampal formation. (A) Eigenvectors of the successor matrix M form spatially periodic fluctuations similar to those observed in the firing rate maps of grid cells recorded in the entorhinal cortex. (B) Spatial autocorrelograms of successor eigenvectors in the two halves of the square and trapezoidal environments. (C) Like real grid cells, the autocorrelograms are more similar in the two halves of the square environment than the two halves of the trapezoidal environment. Figure adapted from Stachenfeld et al. (2017).

While the place and grid cells generated by the successor representation (figures 1.15 & 1.16) may not look overly similar to those observed in the hippocampal formation (figures 1.5 & 1.7), the model is able to capture some characteristics that make it quite appealing. First, due to the independent learning of the successor matrix M and reward weights $\boldsymbol{R}$, latent learning phenomena like that observed by Tolman (Tolman and Honzik, 1930; Tolman, 1948) are intrinsic to the model. Second, since the successor representation is a predictive representation learned through experience, behavioural biases impact the expected future state occupancies causing an experience-dependent skew in the modelled place fields (figure 1.15). This is similar to what has been observed in real place cells (Mehta et al., 2000). Finally, since the model proposes that grid cells encode the transition structure of the environment, changes induced by environmental geometries are reflected in the grid fields (figure 1.16). Consequently the model predicts the reduced grid regularity in trapezoidal environments observed in real grid cells (Krupic et al., 2015).

1.2.5 Summary

In this section we have introduced the field of reinforcement learning and described how the manner in which an artificial agent estimates the value function will impact its action choices in a changing environment. Specifically we introduced model-based and model-free methods, as well as an intermediary approach called the successor representation that forms a predictive map of its environment. We will now investigate the biological validity of these three solution methods by comparing behaviour to human and rat trajectories as they solve a transition revaluation spatial navigation task.

Chapter 2

Transistion revaluation task

2.1 Introduction

Since brain structure is largely conserved across mammals (Finlay and Darlington, 1995), along with the hippocampal representation of space (Ekstrom et al., 2003; Hafting et al., 2005; Ulanovsky and Moss, 2007; Doeller et al., 2010; Yartsev et al., 2011; Yartsev and Ulanovsky, 2013; Maidenbaum et al., 2018), it would be reasonable to question whether this elicits similar spatial behaviour. Furthermore, by comparing this behaviour to artificial agents it might be possible to gain some understanding of *how* the brain utilises its representation of space to facilitate spatial navigation.

As mentioned previously, learning agents utilising model-free, model-based or successor representation methods will all eventually converge to the same set of optimal policies in an unchanging world. However, by introducing systematic changes to the environment it might be possible to pull apart the behaviour of different algorithms. In this chapter we use a transition revaluation task to show that model-free, model-based and successor representation behaviours are dissociable in this spatial navigation framework. Furthermore, we use both likelihood and behavioural similarity analyses to show that the human and rat spatial behaviour is more consistent with an agent implementing a successor representation approach, than either model-based or model-free.

2.2 General framework

We will first cover some general properties of the transition revaluation task, before going into the specific implementations for the rat and human experiments. In all versions of the experiment, the environment consisted of a 10×10 grid of maze modules. These modules could be removed from the grid in order to form impassable barriers in the environment. One of the modules was rewarded and thus was the location of the goal in the maze. Navigation was facilitated by a single distal cue consisting of a black curtain that spanned the majority of one side of the maze (figure 2.2 & 2.9). The goal was kept in the same position with respect to this distal cue throughout all versions of the task. All participants, rats and learning agents were initially trained to navigate to the goal module on the 'open maze', without any maze modules removed. Once trained, they were all put through the same sequence of 25 mazes, with the same sequence of starting locations on each maze (figure 2.1).

The 25 maze configurations were chosen from a sample of 300 as part of a separate study. The aim of that study was to investigate goal-vector cells in the rat hippocampus, and so mazes were initially chosen based on their ability to provide a wide range of distances and angles to the goal, whilst still being possible to solve for rats. The chosen mazes were subsequently ordered in such a way that minimised the spatial correlation between consecutive mazes, thus limiting the knowledge that could be generalised across mazes. Finally, the sequence of starting positions on each maze were chosen to gradually require longer and more tortuous trajectories in order to reach the goal. This was to procure complex trajectories, whilst keeping the rats motivated throughout the task.

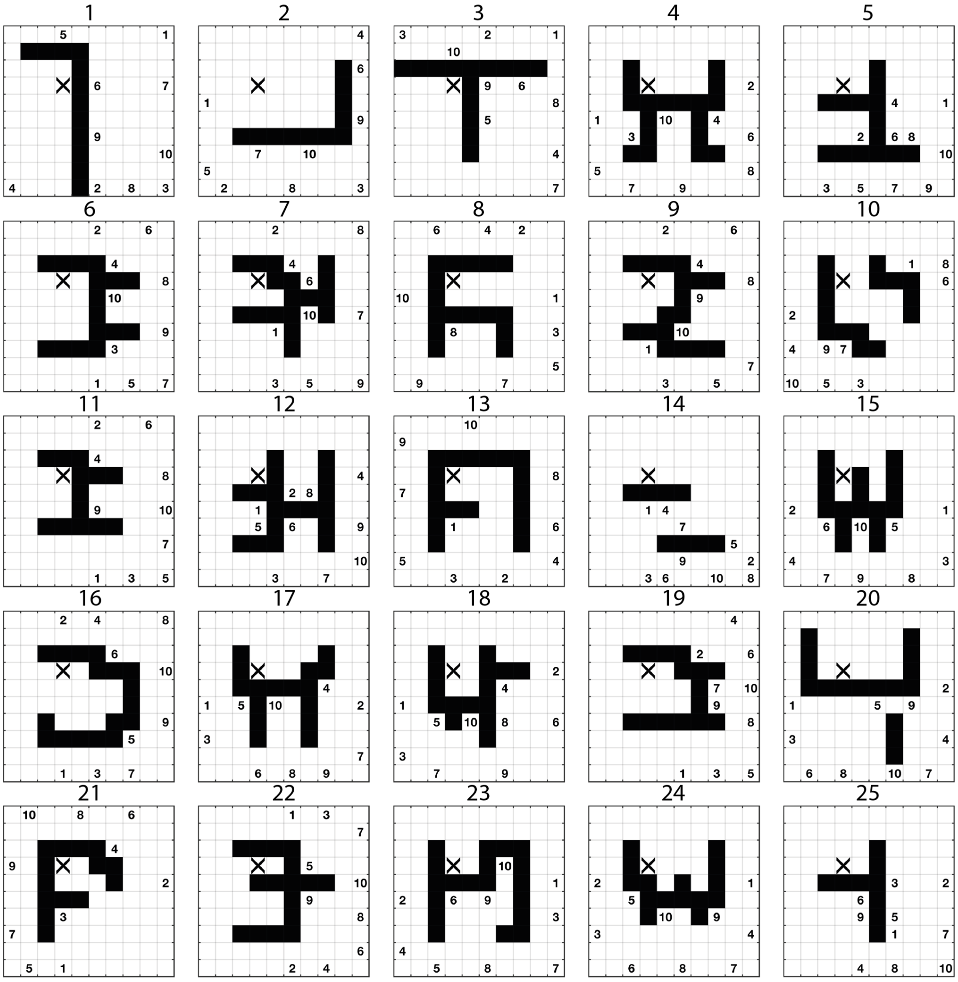


Figure 2.1: Plan view of the maze configurations used in the transition revaluation task. After being trained on the open maze all artificial agents, rats and human participants were put through the same 25 mazes outlined above, with the same sequence of starting positions on each maze. Black squares indicate the maze modules that were removed thus forming impassable barriers in the environment. The numbers 1-10 on each maze configuration indicate the starting positions in order and the 'X' indicates the goal location.

2.3 Reinforcement learning models

The reinforcement learning agents were implemented in a 10 × 10 grid world. At the beginning of the experiment, all agents were endowed with the optimal policy on the 'open maze' to simulate the training phase undertaken by rats and humans. They were then run consecutively on the 25 maze configuration, carrying over all value and model representations between trials. All agent behaviour was simulated using the maximum likelihood parameters fit to the rat/human data, along with an ε-greedy policy where $\varepsilon = 0.1$. This means the agents choose the greedy action 90% of the time and a random action 10% of the time (in order to manage the exploration-exploitation tradeoff). Due to the behavioural variance introduced by this policy, each algorithm was implemented 100 time to produce the distribution of behaviour used for comparison with the rats and humans.

The model-free and successor representation agents were implemented using equations 1.10 and 1.15 at every time step to update their value function. The model-based agent updated its model at every time step by observing the states adjacent to its current state and implementing equation 1.11. It would then sample a map from the model and use it to plan the shortest route to the goal from its current position (shortest routes were calculated using the Hart et al. (1968) A-star search algorithm). In the event of multiple equally short routes to the goal, their respective actions were sampled with equal probability.

2.4 Rat experiment

2.4.1 Methods

2.4.1.1 Animals

Nine adult male Lister Hooded rats were handled daily (at start of training: 10-20 weeks old, 350-400 g) and housed communally in groups of three. All rats were subjected to a reverse light-dark cycle (11:11 light:dark, with 1 hour ×2 simulated dawn/dusk) and were on food-restriction sufficient to maintain 90% of free-feeding weight, with ad libitum access to water. The free-feeding weight was continuously adjusted according to a calculated growth curve for Lister Hooded Rats (Clemens et al., 2014). Six rats were naïve, while three rats had previously been trained for 2-3 weeks in a shortcut navigation task for a different maze setup. The procedures were conducted according to UCL ethical guidelines and licensed by the UK Home Office subject to the restrictions and provisions contained in the Animals Scientific Procedures Act of 1986.

2.4.1.2 Protocol

All procedures were conducted during the animals' dark period. The experiment was carried out in a custom-made modular 2x2m square maze composed of 100 identical square platform tiles elevated 50cm above the ground via two pieces of wood supports fit together through their long slits (figure 2.2). The maze was constructed from Medium Density Fibrewood, with the platforms painted in grey. Each platform contained a plastic well (32mm diameter, 9mm depth) at its centre, which could be attached to polymeric tubing system installed beneath the maze. This tubing allowed the experimenter to reward the rat at the goal module by soundlessly filling the well with chocolate milk (0.1 ml). The maze was surrounded on all sides by a white curtain, with a black sheet overlaid on one side to provide a single extra-maze cue (figure 2.2). The goal module was always in the same position with respect to this cue (figure 2.1).

An initial familiarisation phase lasted for three days. During the first day, the rats received a small amount (0.1ml per rat) of chocolate milk in the home cage to decrease neophobia on the maze. For the subsequent two days, each rat underwent two 15 minute maze familiarisation sessions, in which the rat was placed at the centre of the maze and would forage for pieces of chocolate cereal (Weetos) scattered throughout the maze. More cereal was concentrated in the centre to encourage the animal to be comfortable in the middle of the maze. The experimenter was present beside the maze inside the curtained area throughout the session, and between sessions the black sheet was rotated 90° counter-clockwise.

After the familiarisation phase, rats began to be trained to navigate to the goal location. In each training trial the rat had 45s to find the goal module, during which

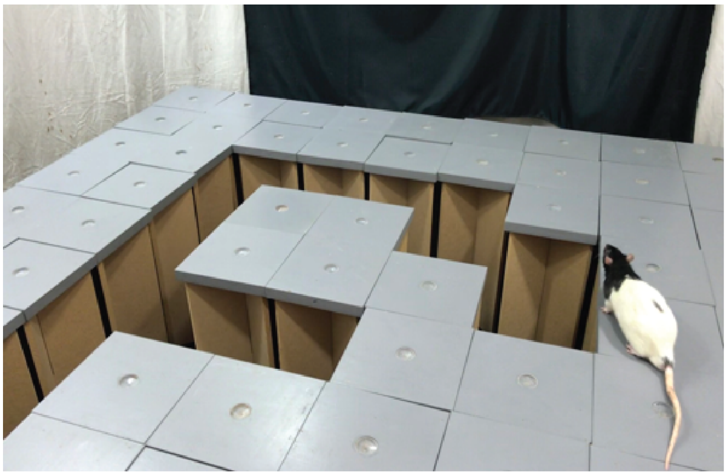


Figure 2.2: The maze environment used for the rat experiment. The environment consists of 100 removable mazes modules with a black curtain over one of the surrounding edges to provide a single extra-maze cue. Reward can be dispensed at the goal module by filling the well with chocolate milk via polymeric tubing beneath the maze.

the experimenter stood motionless next to the maze. Training consisted of three stages. The first stage lasted one day and consisted of two 15 minute sessions, during which the goal module's well was filled with 0.1ml of chocolate milk and the rats were initially placed on the modules adjacent to the goal. For each subsequent training trial, the rat's starting position would be shifted one module anticlockwise. If the rat made two consecutive direct runs to the goal (without exploration of other parts of the maze), the next trial begun one module further away from the goal. Conversely, if the rat failed two consecutive training trials, the next trial begun one module closer to the goal until the rat was back at the goal-adjacent modules. In this first stage, the rats were always placed facing the goal. The second training stage followed the same procedure was as stage one, but the number of trials was fixed to 16. This procedure was followed every day until the rat was able to make direct runs from the edges of the maze. The third and final training stage was also similar to stages one and two, except the number of daily trials could be increased up to 25. Furthermore, the rat's starting position and orientation was randomised and a delay in the release of chocolate milk was introduced. This delay started at 1s and was gradually increased until the rat could wait at the goal location for 5s before the chocolate milk was released. This procedure was followed until the rats were able to successfully navigate directly to the goal and on at least 90% of trials. The training phase took on average 24 sessions

Following the training phase, rats were run on the 25 maze configurations with the starting positions indicated in figure 2.1. Trials were 45s long and rats were required to navigate to the goal within this time and wait for 5s in order to receive the reward (0.1ml of chocolate milk). If the rat failed to reach the goal, it moved onto the next trial. At the beginning of each day, rats were given a brief ‘reminder session’ that consisted of 5 trials from phase 3 of the training phase, and then the rats would usually go on to complete 3 configurations per day.

2.4.2 Results

The rat trajectories from the testing phase were smoothed and discretised onto the 10×10 modular grid for comparison with the reinforcement learning agents. Only transitions between states were included in these trajectories. Rats displayed a steady increase in their ability to navigate to the goal with greater exposure to a maze configuration (proportion goal reached during first 5 trials vs last 5 trials, mean ± sd: 0.80 ± 0.10 vs 0.90 ± 0.16; $t(8) = -3.95$, $p < 0.01$; Figure 2.3A), along with an increased efficiency in the routes used to reach the goal (deviation from optimal path first 5 trials vs last 5 trials: 10.2 ± 1.6 vs 7.7 ± 0.9; $t(8) = 4.02$, $p < 0.01$; Figure 2.3B).

This data was next used to estimate the maximum likelihood parameter values for the learning rates α and discount factors γ used in the update equations 1.10, 1.11 and 1.15 for the reinforcement learning algorithms (table 2.1). To calculate the likelihoods of the models, the rat behaviour was input into each agent and at every time step the agent’s value function V for potential next states s_i was passed through a softmax function $\frac{e^{V(s_i)}}{\sum_i e^{V(s_i)}}$ to generate a probability distribution over the possible actions. The probability corresponding to the action made by the rat was then used in the likelihood function and the agent moved onto the next time step. This process

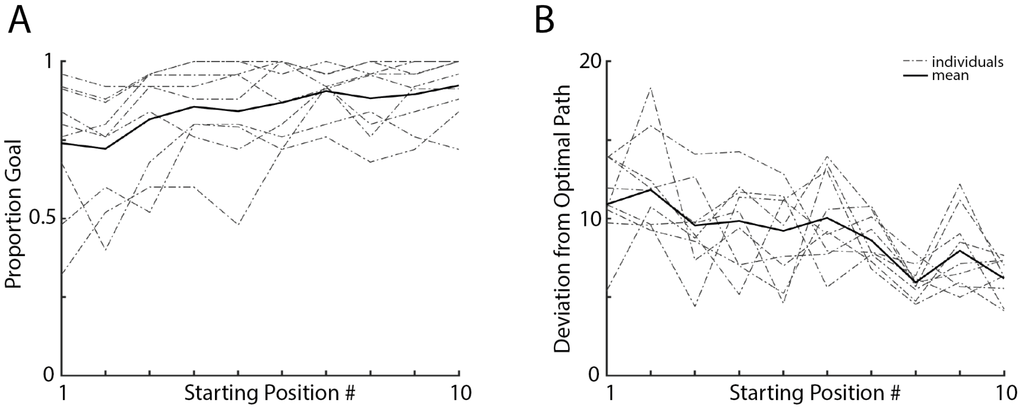


Figure 2.3: Rats improved at navigating to the goal with increased exposure to a new maze configuration. (A) Rats were able to more consistently navigate to the goal towards the later trials of a new maze configuration (proportion goal reached during first 5 trials vs last 5 trials, mean ± sd: 0.80 ± 0.10 vs 0.90 ± 0.16; $t(8) = -3.95$, $p < 0.01$). (B) The trajectories used in these later trials were also more efficient (deviation from optimal path first 5 trials vs last 5 trials: 10.2 ± 1.6 vs 7.7 ± 0.9; $t(8) = 4.02$, $p < 0.01$).

was repeated and combined across all animals and the parameters maximising each agent's likelihood function were obtained using the *fmincon* constrained optimisation function in MATLAB 2018b. Since the model-free agent primarily operates using the action-value function Q, the value function V of potential next states was computed following equation 1.9 and a greedy policy. Interestingly, the inferred discount parameters γ were somewhat similar for all types of agent. This suggests all algorithms are indeed trying to approximate the same value function (equation 1.6). It is also important to note that the $\alpha = 1$ for the model-based algorithm means that any stochasticity in the sampling of a map from the agent's model is removed - every observation becomes part of its model with 100% certainty.

Model	α	γ
Model-free	0.22	0.68
Model-based	1	0.79
Successor representation	0.81	0.79

Table 2.1: Maximum likelihood parameters for all reinforcement learning models based on the rat behaviour. α is the learning rate and γ is the discounting parameter.

The maximum likelihood estimates for these parameters (figure 2.4A) suggests that the successor representation provides the best match to the rat behaviour out of the three models (Likelihood Ratio Test: successor representation vs. model-free, $LR = 291.9$; successor representation vs. model-based, $LR = 450.3$). On average, the quality of this fit was slightly worse in the last 5 trials of each configuration (figure 2.4B). Furthermore, the model fits appear to systematically vary in quality for specific configurations, with less variation between models (figure 2.4C). This configuration variability did not correlate with the rats' goal-reaching on the mazes (Pearson's correlation: model-free, $\rho = -0.13$, $p = 0.55$; model-based: $\rho = -0.15$, $p = 0.49$; successor representation: $\rho = -0.11$, $p = 0.60$).

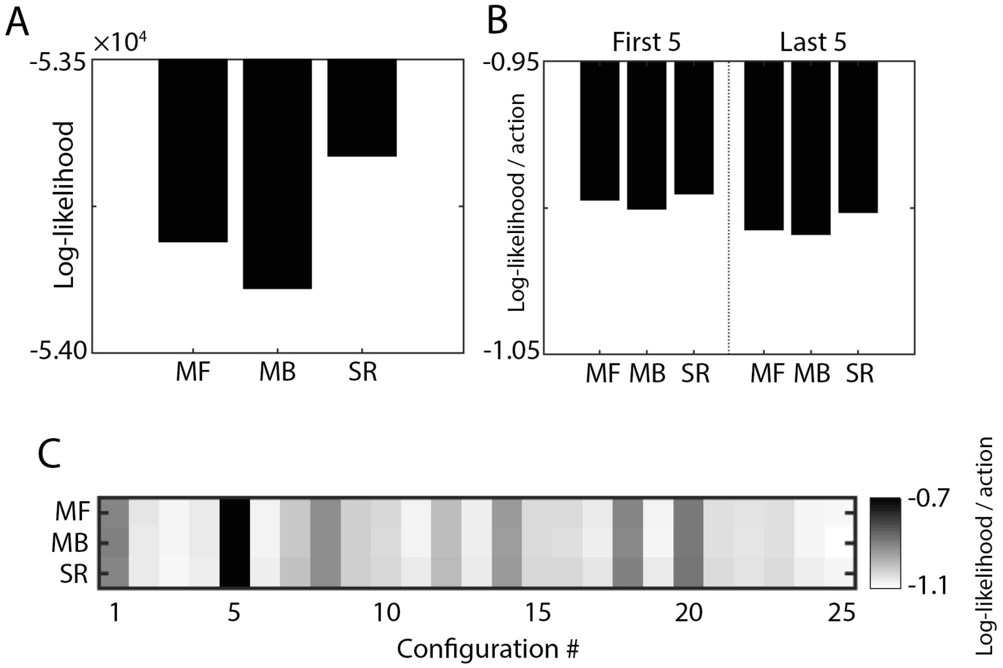


Figure 2.4: Maximum log-likelihood of the reinforcement learners fitted to the rat behaviour. (A) Maximum log-likelihoods of each of the models. (B) Average log-likelihood per action from the first 5 trials and last 5 trials on a configuration. (C) Average log-likelihood per action for each configuration.

To investigate whether these differences in likelihoods actually transfer into meaningful behaviours, we used the maximum likelihood model parameters to simulate trajectories for each reinforcement learner ($n = 100$) using an ε-greedy policy ($\varepsilon = 0.1$). The number of time steps that the learners had to reach the goal was set to the mean length of the rat trajectories on trials when they did not reach the reward (mean ± sd: 29 ± 17 time steps).

Examining the proportion of trials in which the rats and reinforcement learners were able to navigate to the goal revealed substantial maze-dependent variations in goal-reaching (figure 2.5A). Using this to rank the maze configurations in order of relative difficulty for each agent, there was only a significant correlation between the successor representation and rat difficulty rankings (Spearman's rank correlation: $\rho = 0.41$, $p = 0.04$; figure 2.5B). Furthermore, viewing how the proportion of goal-reaching varied with exposure to a new configuration (figure 2.5C) revealed a slight increase for the model-based algorithm (first 5 trials vs last 5 trials: $t(8) = 2.6$, $p = 0.03$), no significant change for the successor representation ($t(8) = 0.8$, $p = 0.43$), and a strong decrease in goal-reaching for the model-free agents ($t(8) = 4.1$, $p < 0.01$). This decrease in goal-reaching for the model-free agents is due to the increasing difficulty of starting positions on a maze, combined with it being unable to adapt quickly enough to the changing environment. This is in spite of it being the second most likely model (figure 2.4A) and exemplifies the inflexibility of model-free frameworks as well as the disassociation between likelihoods and behaviour in sparse, real-world tasks. Importantly, none of the algorithms were able to recapitulate the rats' goal-reaching performance on the maze.

Whilst goal-reaching provides a useful summary of how animals and agents generally compared throughout the task, it does not provide any bearing on the similarity of the actual behaviour used. To assess this, we next quantified every trajectory by calculating 3 measures on them; *rotational velocity*, *diffusivity* and *tortuosity*. The rotational velocity is a rolling measure of the angle between consecutive points in a trajectory separated by δ time steps. This is then normalised by δ and the mean was taken across the entirety of a trajectory (equation 2.1). Similarly, the diffusivity is a rolling measure of the squared Euclidean distance travelled between consecutive points in a trajectory separated by δ time steps. Again, this was then normalised by δ and the mean was taken the trajectory (equation 2.2). Finally the tortuosity is a measure of the bendiness of a trajectory and is equal to the total path distance travelled divided by the Euclidean distance travelled (equation 2.3).

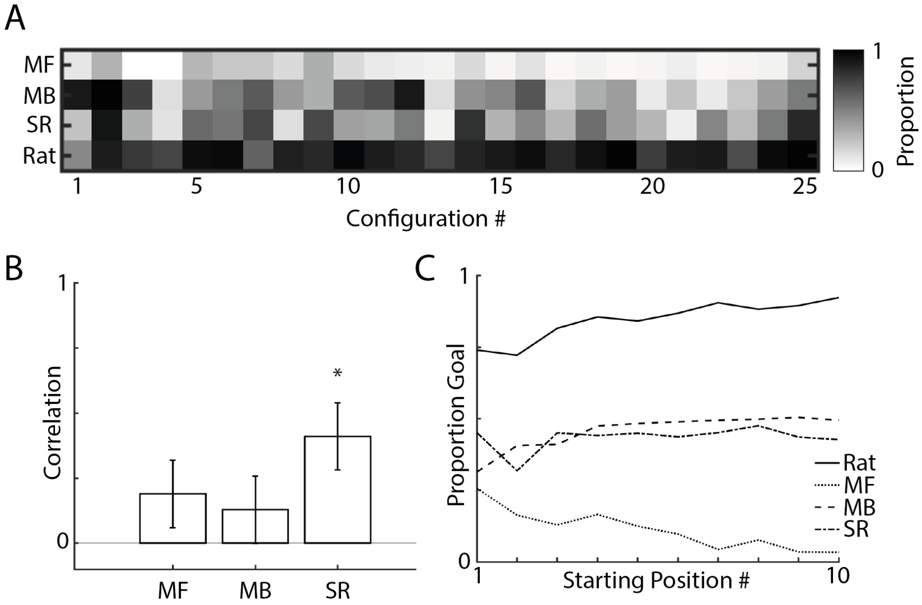


Figure 2.5: Rat and reinforcement learner goal-reaching. (A) Proportion of trials on a maze configuration where the simulations/rats successfully navigated to the goal. (B) Spearman's rank correlation between the maze configurations that the rats found most easy/difficult and the those that the reinforcement learners found most easy/difficult. (C) Proportion of successful goal-reaching with exposure to the maze configuration.

$$\text{Mean rotational velocity} = \sum_{t=1}^{T-\delta} \frac{\arctan 2(x_{t+\delta} - x_t, y_{t+\delta} - y_t)}{\delta(T-\delta)} \tag{2.1}$$

$$\text{Mean diffusivity} = \sum_{t=1}^{T-\delta} \frac{(x_{t+\delta} - x_t)^2 + (y_{t+\delta} - y_t)^2}{\delta(T-\delta)} \tag{2.2}$$

$$\text{Tortuosity} = \frac{T}{\sqrt{x_T - x_1)^2 + (y_T - y_1)^2}} \tag{2.3}$$

For both the rotational velocity and diffusivity measurement, $\delta = 3$. In the event that the denominator of equation 2.3 is zero (i.e. the trajectory starts and finishes in the same place), then the denominator was set to 0.5 to keep the measure smooth and finite. The rotational velocity and diffusivity measures were chosen to describe translationally invariant, local movement information, whereas the tortuosity measure provides a more high level description of trajectory complexity.

By using dimensionality reduction techniques such as t-distributed stochastic neighbour embedding (t-SNE) (van der Maaten and Hinton, 2008) which minimises the difference in the distribution of pairwise distances within the actual data and its 2-dimensional embedding, it is possible to visualise the behavioural space averaged over the course of the experiment (figure 2.6A). Consequently, we can see that the agent and rat behaviours form fairly distinct clusters. To investigate this further, we measured the Mahalanobis distance between the 3 trajectory measures to calculate the dissimilarity matrix of the trajectories for every starting position across every maze. In particular, the Mahalanobis distance was used because it takes the covariance between these measures into account when calculating similarity. The average of the dissimilarity matrices (figure 2.6B) revealed that the rat trajectories are significantly closer to the successor representation trajectories in this behavioural space (successor representation vs. model-based: $t(16) = 9.3$, $p < 0.001$; successor representation vs. model-free: $t(16) = 11.0$, $p < 0.001$; figure 2.6C).

To make use of this clustering of agent behaviour, we next built a k-nearest neighbours (kNN) classifier for every starting position on every configuration, and used it to decode agent identity from trajectory data. With the number of neighbours $k = 100$, the kNN classifier was able to decode the underlying reinforcement learning algorithm well above chance according to a 10-fold cross-validation (figure 2.7). Despite being the best agent at navigating to the goal (figure 2.5) the model-based algorithm was the most difficult to decode, particularly on mazes where the structure imposes limits on the trajectory measures being used for classification (e.g. tortuosity). There was no remarkable correlation between goal-reaching and decoder performance on mazes (Spearman's Rank correlation: model-free, $\rho = 0.27$, $p = 0.20$; model-based, $\rho = 0.19$, $p = 0.36$; successor representation $\rho = -0.06$, $p = 0.78$)

Finally, we used the rat trajectory data as input to the k-NN classifier in order to see

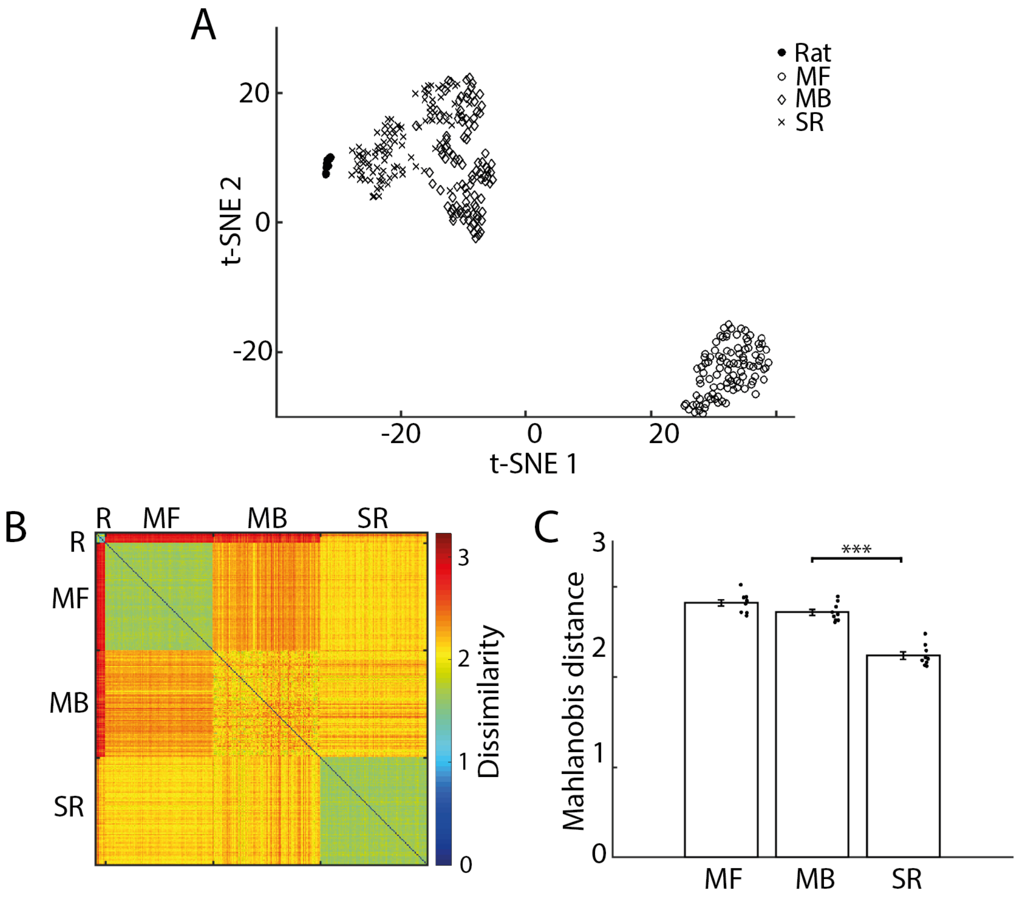


Figure 2.6: Clustering of navigational behaviour during the rat experiment. Three measures were calculated on each trajectory to quantify the trajectories: tortuosity, average rotational velocity and average diffusivity. (A) Using t-SNE (van der Maaten and Hinton, 2008) to view the average of this behaviour throughout the experiment reveals distinct clusters. (B) Using these measures to calculate the Mahalanobis distance between rats and the reinforcement learners for every trial allows us to construct the average behavioural dissimilarity matrix. (C) The average Mahalanobis distance between each of the reinforcement learning algorithms and the rat behaviour during the experiment.

what reinforcement learning agent was most often predicted. Consistent with the likelihood and behavioural similarity analyses, rat behaviour was most commonly mistaken as an agent employing a successor representation algorithm (figure 2.8A). Furthermore, rat trajectories were more likely to mistaken as a successor representation agent towards the later trials on a maze configuration (first 5 trials vs. last 5 trial: $t(8) = 4.3$, $p < 0.01$; (figure 2.8B)). Lastly, the extent of these predictions did not correlate with the proportion of goal-reaching observed by successor represen-

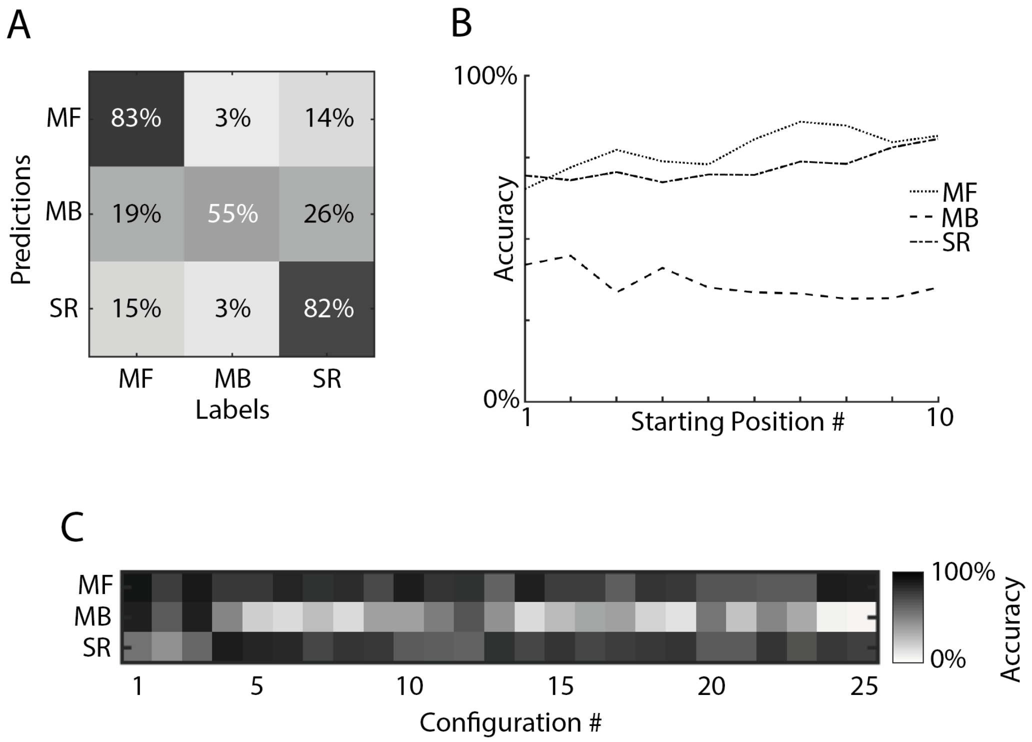


Figure 2.7: Using k-nearest neighbours (k-NN) to classify reinforcement learners in the behavioural space. (A) Confusion matrix for the k-NN classification from a 10-fold cross-validation. (B) How ability to decode reinforcement learning algorithm changes with exposure to the maze. (C) How ability to decode reinforcement learning algorithm varies with individual maze configurations.

tation agents on specific mazes (Spearman's rank correlation: $\rho = -0.13$, $p = 0.54$; figure 2.8C).

2.5 Human experiment

2.5.1 Methods

For the human version of the task, 18 healthy participants (9 female; aged = 24.6 ± 5.9, mean ± sd) were recruited from the UCL Psychology Subject Pool and trained to navigate to an unmarked goal in virtual arena of approximately the same relative proportion as for the rats. All participants gave written consent to participate in the study in accordance with the UCL Research Ethics Committee. Participants were reimbursed for their time as well as a bonus of up to £25 for good performance

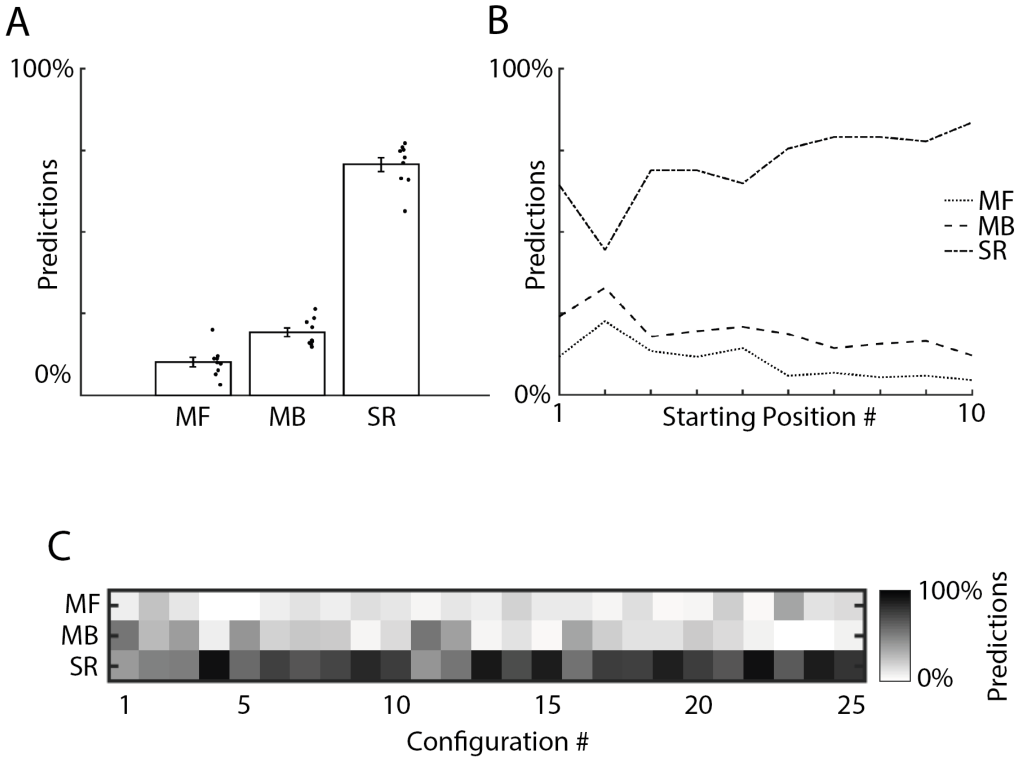


Figure 2.8: Using the k-nearest neighbours (k-NN) classifier to classify rat behaviour as a reinforcement learning algorithm. (A) The proportion of predictions made by the k-NN classifier when shown the rat behaviour. (B) How these predictions varied with the rats' exposure to the maze configuration. (C) How these predictions varied across maze configurations.

in the testing phase. Participants experienced the virtual environment via a HTC Vive virtual reality headset whilst sat on a swivel chair. They were able to adjust movement speed using the HTC Vive controller and movement direction was controlled by the participant's orientation on the chair. Upon successful navigation to the goal module, participants were informed of their financial reward along with the presence of a revolving gold star at the goal location. In accordance with the rodent experiment, navigation was aided by the presence of a black distal cue that took up the majority of one of the walls (figure 2.9). Goal location, maze configurations and starting positions were all defined with respect to this distal cue and were identical to the rodent experiment. Importantly, a fog lined the floor of the maze (figure 2.9) to prevent the participants from understanding what maze modules were missing until they were at adjacent locations. This also provided a better match to visual information available to the rats - which are known to have less visual acuity and binocular depth perception (Heffner and Heffner, 1992). Seamless textures were applied to the floor and walls of the virtual environment, and these were rotated every 10 trials to prevent them from being used as extraneous cues for navigation.

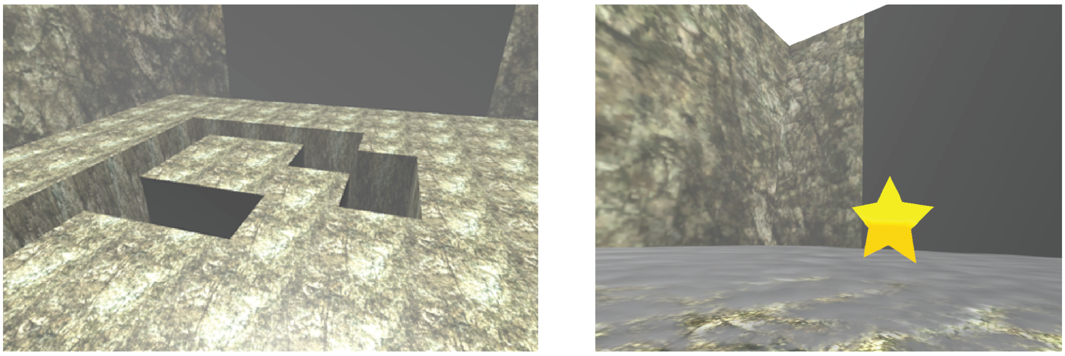


Figure 2.9: The virtual environment used for the human experiment. The environment had the same proportions as the rat environment and consisted of 100 removable mazes modules with a black curtain over one of the surrounding edges to provide a single extra-maze cue. A seamless texture was applied to the maze modules and walls and a fog lined the floor of the maze (see right image) to ensure humans had to rely on spatial memory to understand the maze structure. Reward was indicated by a gold star that would appear at the goal module when the participant successfully navigated to it.

The experiment took place over four sessions on four consecutive days. The majority of the first session was usually spent training the participants to navigate to the goal module. To accelerate this learning process, the participants were initially able see a revolving gold star in the goal location. As they progressed through the training session the star became increasingly transparent until invisible, with the star only appearing again upon successful navigation to the goal module. Along with the decreasing visibility of the goal, the participants' starting positions were moved progressively further from the goal in a similar manner to the rat training phase. All training and testing trials were 45s in length. Training was terminated when the participants were able navigate to the hidden goal on at least 80% of trials after being randomly placed at the far edges of the environment. Mean time to complete this training was 41 ± 21 minutes. In order to make the participants' experience similar to that of the rodents, they were not given any explicit information about the nature of the task - only that financial reward was hidden in the environment in the form of a gold star and their task was to maximise their financial return as quickly and efficiently as possible.

The testing took place over the remaining sessions and on average lasted 125 ± 25 minutes, with participants encouraged to take short breaks every 10-20 trials to reduce virtual reality sickness. At the beginning of each testing session, participants completed a short ‘reminder task’, which consisted of 5 trials from the end of the training phase.

2.5.2 Results

To compare the human behaviour to that of the reinforcement learning agents, participants’ trajectories were discretised into the transitions made on the underlaying 10×10 modular grid. Similarly to the rats, participants also displayed a steady increase in their ability to navigate to the goal with increased exposure to a maze configuration (proportion goal reached during first 5 trials vs last 5 trials: 0.93 ± 0.06 vs 0.97 ± 0.05; $t(17) = -6.35$, $p < 0.001$; Figure 2.10A), along with an increased efficiency in the routes used to reach the goal (deviation from optimal path first 5 trials vs last 5 trials: 3.95 ± 2.27 vs 2.72 ± 1.54; $t(17) = 4.97$, $p < 0.001$; Figure 2.10B).

Just like in the rat analysis, these trajectories were used to estimate the maximum likelihood parameters for each of the models (figure 2.2).

Model	α	γ
Model-free	0.45	0.75
Model-based	1	0.81
Successor representation	0.93	0.79

Table 2.2: Summary of the maximum likelihood model parameters fitted to the human behaviour. α is the learning rate and γ is the discounting parameter.

Interestingly, the discount parameters γ were similar to those inferred from the rat data (0.68, 0.79 and 0.79 for model-free, model-based and successor representation respectively), while the inferred learning rates α were generally larger for the

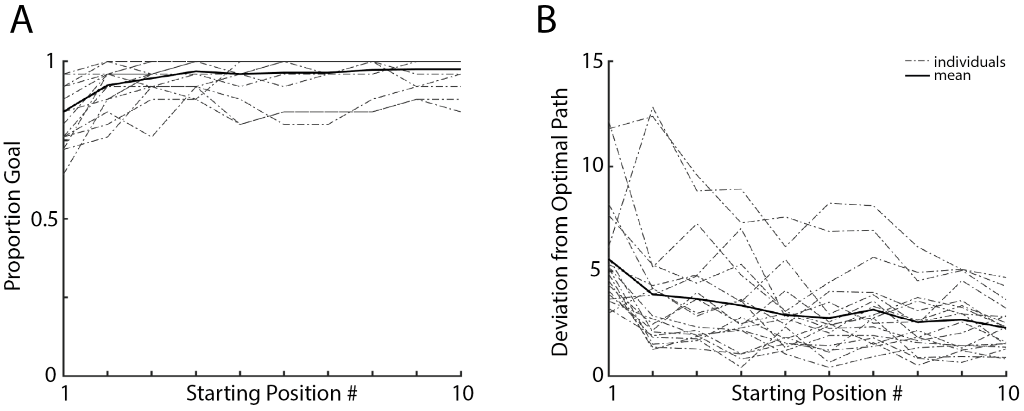


Figure 2.10: Human participants improved at navigating to the goal with increased exposure to a new maze configuration. (A) Participants were able to more consistently navigate to the goal towards the later trials of a new maze configuration (first 5 trials vs last 5 trials: 0.93 ± 0.06 vs 0.97 ± 0.05; $t(17) = -6.35$, $p < 0.001$). (B) The trajectories they used in these later trials were also more efficient (deviation from optimal path first 5 trials vs last 5 trials: 3.95 ± 2.27 vs 2.72 ± 1.54; $t(17) = 4.97$, $p < 0.001$).

human participants. Furthermore, the successor representation was again the most likely model to explain the biological data (figure 2.11A; Likelihood Ratio Test: successor representation vs. model-free, $LR = 150.3$; successor representation vs. model-based, $LR > 500$). As with the rat likelihoods, all the reinforcement learning algorithms appeared to systematically provide a better or worse fit to the human data depending on the maze configuration (figure 2.11C), with a significant correlation between the human and rat model fits on these mazes (Spearman's correlation: model-free, $\rho = 0.50$, $p < 0.01$; model-based, $\rho = 0.59$, $p < 0.01$; successor representation, $\rho = 0.54$, $p < 0.01$). Additionally, the reinforcement learning algorithms seemed to systematically provide a better fit to the trials at the beginning of a new maze (figure 2.11B).

Once again, in order evaluate whether these differences in model likelihoods actually relate to meaningful similarities in behaviour, we next simulated each artificial agent $n = 100$ times with the maximum likelihood parameter estimates from table 2.2. Each algorithm implemented an ε-greedy policy with $\varepsilon = 0.1$. Based on the length of the trajectories of human participants when they did not reach the goal (mean ± sd: 42 ± 8), the artificial agents were given a limit of 42 time steps.

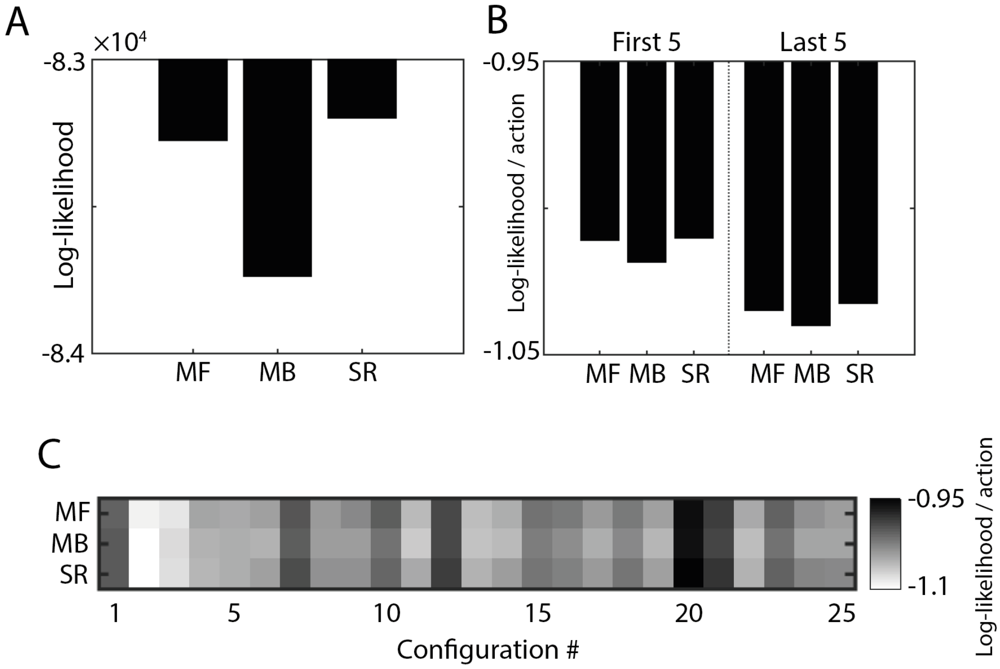


Figure 2.11: Maximum log-likelihood of the reinforcement learners fitted to the human behaviour. (A) Maximum log-likelihoods of each of the models. (B) Average log-likelihood per action from the first 5 trials and last 5 trials on a configuration. (C) Average log-likelihood per action for each configuration.

Using the proportion of goal-reaching trials to rank the mazes in terms of relative difficulty (figure 2.12A), there was only a significant correlation between the maze difficulty rankings for the humans and the successor representation agents (Spearman's rank correlation: $\rho = 0.42$, $p = 0.04$; figure 2.12B). There was also a significant correlation between goal-reaching on configurations for humans and rats ($\rho = 0.52$, $p < 0.01$). Once again, despite being the second most likely algorithm from the likelihood analysis, the model-free agent was unable to transfer this into goal-directed behaviour and on average got progressively worse throughout configurations (first 5 trials vs. last 5 trials: $t(8) = -3.0$, $p = 0.02$; figure 2.12C). Since the successor representation agent is slower to converge on the optimal policy than the model-based agent, it is able to outperform the model-based algorithm on mazes where the optimal policy anticorrelates with that of the preceding maze (e.g. mazes 9, 14, 22 - see figure 2.1 for maze layouts). Both model-based and successor representation agents progressively improved within a configuration (first 5 trials vs. last 5 trials: model-based, $t(8) = 2.8$, $p = 0.02$; successor representation, $t(8) = 2.8$,

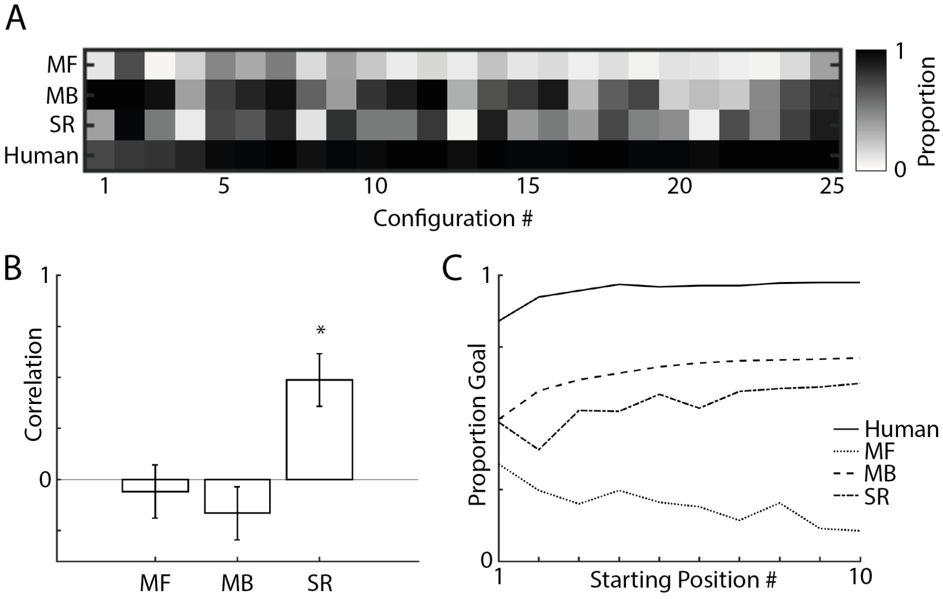


Figure 2.12: Human participant and reinforcement learner goal-reaching. (A) Proportion of trials on a maze configuration where the simulations/rats successfully navigated to the goal. (B) Spearman's rank correlation between the maze configurations that the participants found most easy/difficult and the those that the reinforcement learners found most easy/difficult. (C) Proportion of successful goal-reaching with exposure to the maze configuration.

$p = 0.03$), with the model-based agents being closer to the human participants in terms of absolute performance (figure 2.12C).

In order to investigate the behavioural similarities between agents and participant trajectories, we followed the method from the rat analyses by quantifying every trajectory using three measures: mean rotational velocity, mean diffusivity and tortuosity. Viewing the 2-dimensional embedding of this behaviour space using t-SNE (van der Maaten and Hinton, 2008) reveals clustering of trajectories (figure 2.13A). In this embedding, there appears to be more of an overlap between model-based and successor representation trajectories, something that is more evident when observing the dissimilarity matrix averaged across all trials (figure 2.13B). Using this dissimilarity matrix to calculate the average Mahalanobis distance between the agents and humans (figure 2.13C), we see a small but significant similarity between the human behaviour and the successor representation agents, compared to the model-based agents ($t(34) = 2.3, p = 0.02$).

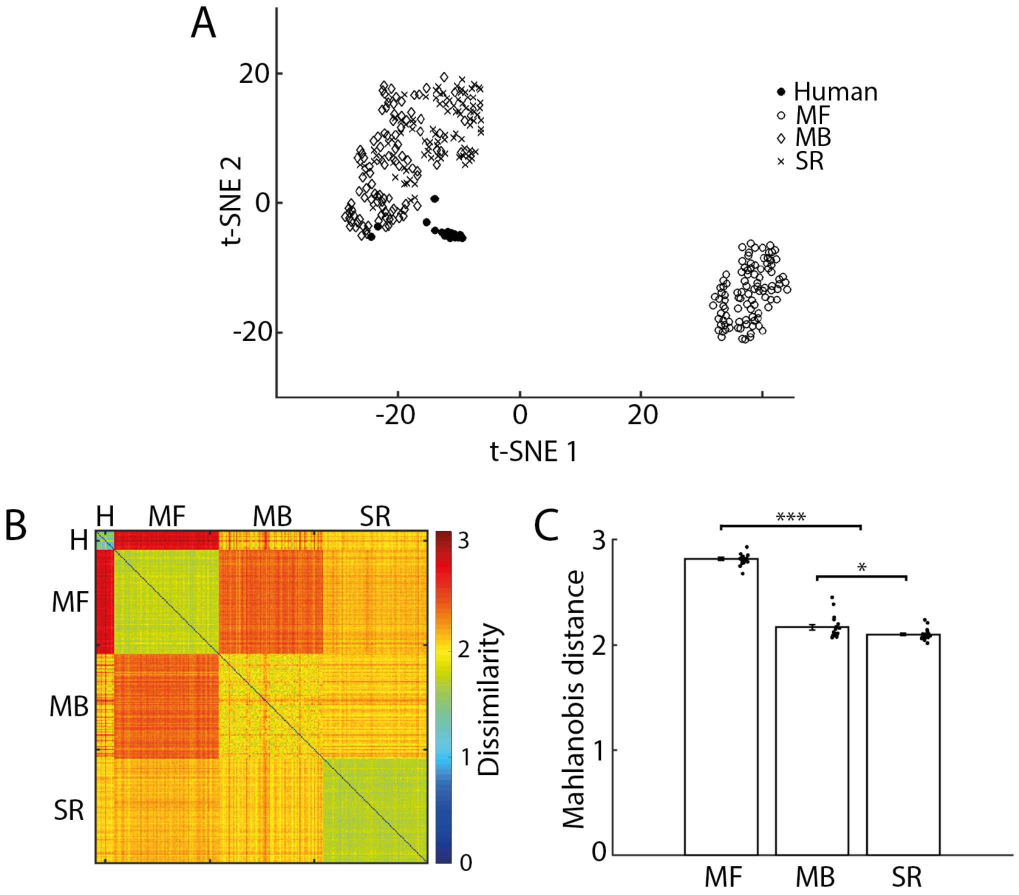


Figure 2.13: Clustering of navigational behaviour during the human experiment. Three measures were calculated on each trajectory to quantify the trajectories: tortuosity, average rotational velocity and average diffusivity. (A) Using t-SNE (van der Maaten and Hinton, 2008) to view the average of this behaviour throughout the experiment reveals distinct clusters. (B) Using these measures to calculate the Mahalanobis distance between human participants and the reinforcement learners for every trial allows us to construct the average behavioural dissimilarity matrix. (C) The average Mahalanobis distance between each of the reinforcement learning algorithms and the human behaviour during the experiment.

To test the reliability of these differences, we trained a k-NN classifier $(k = 100)$ to decode agent identity based on the three trajectory measures. Using a 10-fold cross-validation, the classifier was able to successfully decode the agent trajectories well above chance level (figure 2.14). As with the previous classifier, the model-based behaviour appeared the most difficult to decode.

Lastly, using the human participant data as input to the k-NN classifier predomi-

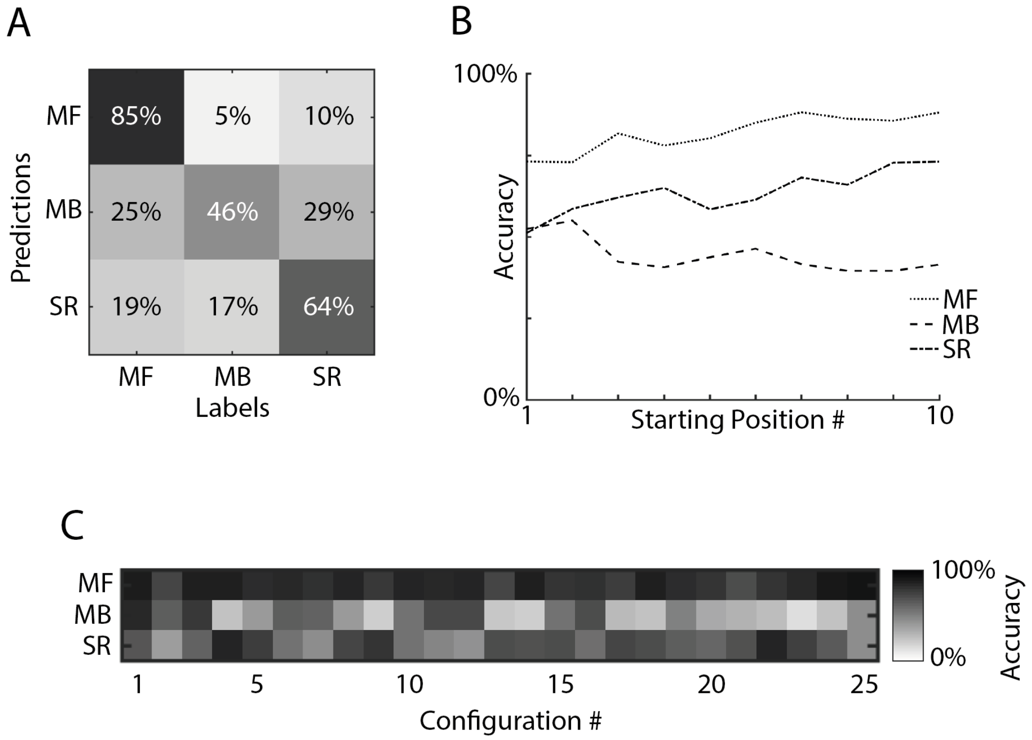


Figure 2.14: Using k-nearest neighbours (k-NN) to classify reinforcement learners in the human behavioural space. (A) Confusion matrix for the k-NN classification from a 10-fold cross-validation. (B) How ability to decode reinforcement learning algorithm changes with exposure to the maze. (C) How ability to decode reinforcement learning algorithm varies with individual maze configurations.

nantly yielded predictions of successor representation agents (figure 2.15A). This observation was largely driven by the later trajectories on a configuration (successor representation vs. model-based predictions: first 5 trials, $t(8) = 0.22$, $p = 0.83$; last 5 trials, $t(8) = 4.1$, $p < 0.01$; figure 2.15B), which was when the humans were best at reaching the goal (figure 2.12C).

2.6 Discussion

Here, we used a transition revaluation task to investigate the spatial navigational strategies of rats and human participants by comparing them to model-free, model-based and successor representation reinforcement learning agents. Using both likelihood and behavioural similarity analyses, we show that the biological behaviour

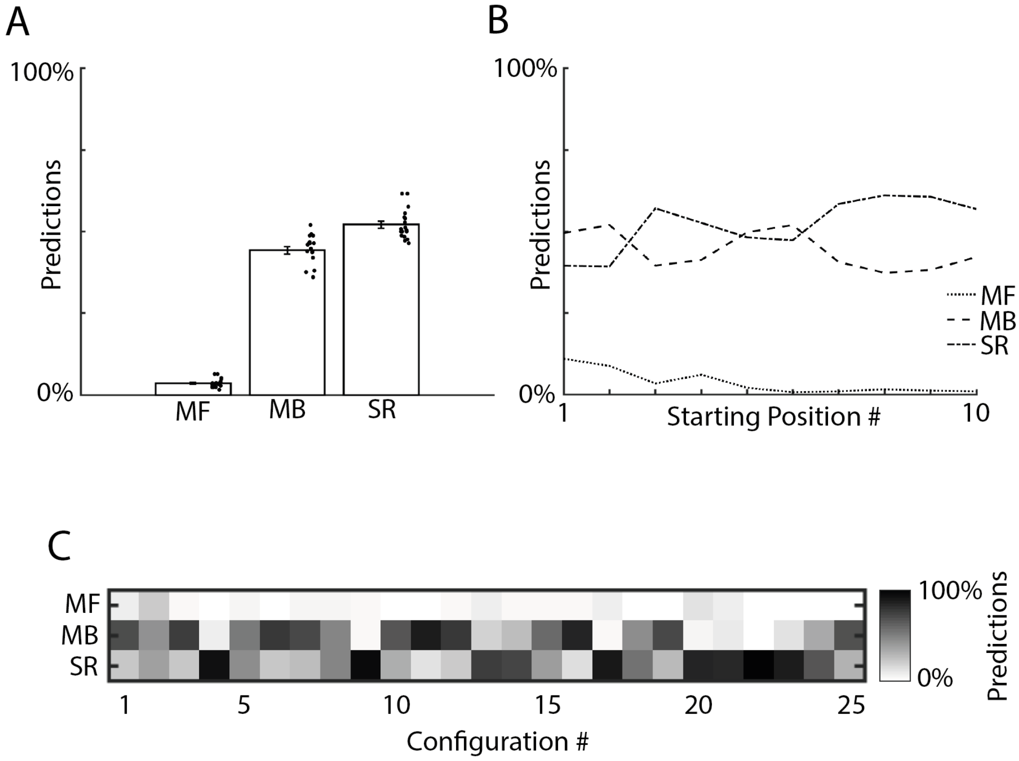


Figure 2.15: Using the k-nearest neighbours (k-NN) classifier to classify human behaviour as a reinforcement learning algorithm. (A) The proportion of predictions made by the k-NN classifier when shown the rat behaviour. (B) How these predictions varied with the human participants' exposure to the maze configuration. (C) How these predictions varied across maze configurations.

is most consistent with an artificial agent implementing a predictive map of its environment. This was further evidenced by a k-nearest neighbours classifier of reinforcement learners predominantly predicting the human and rat trajectories as successor representation agents.

During the likelihood analysis, there was a high maze-to-maze variability in the model fits, and a strong overlap between the quality of these fits on mazes for both the rat and human behaviour. In fact the rat and human behaviour share 6 out of their 10 best fitting mazes (configurations: 1, 7, 10, 12, 18, 20), and 7 of the their 10 worst fitting mazes (configurations: 2, 3, 4, 6, 11, 13, 19). Looking at these more closely, it is notable that the majority of the worst fitting mazes have very little overlap in optimal policy with the mazes that immediately precede them. Furthermore, half of the best fitting mazes (configurations: 7, 12, 18) have a very strong overlap in optimal policy with the immediately preceding maze. This suggests that when the previous policy can be generalised to the new maze layout, the reinforcement learning models are better at fitting the biological behaviour. Indeed, these algorithms are built upon a framework of exploiting knowledge to maximise reward, and the reciprocal of this means that when there is little knowledge to exploit the agents do not have a mechanism for logical and directed exploration.

All of the maximum likelihood γ parameters were similar for both the rat and human behaviour, suggesting that they approximate similar value functions. This also appears to be relatively conserved across species on this task. However, the human behaviour was indicative of algorithms with higher learning rates than for the rats, suggesting they accrue and exploit new knowledge faster.

Despite similarities in some of the goal-reaching capabilities, summarising trajectory data using the rotational velocity, diffusivity and tortuosity robustly showed distinct clustering of behaviour for each reinforcement learning strategy. By utilising this clustering, we could calculate the similarity between the simulated data and the biological data. For both the rat and human data, trajectories summarised in the behaviour space were more similar to the successor representation agents than the other models. This effect was most pronounced in the rat experiment, with the human behaviour also being somewhat similar to the model-based agent.

Finally, to take advantage of the robust clustering of behaviour, we trained a k-nearest neighbours classifier to decode agent identity from the trajectory data. The decoder was able to classify well above chance level. Consistent with both the likelihood and behavioural similarity analyses, the classifier most often confused the rat and human trajectories with those of a successor representation agent. This effect was more pronounced in the later trials on a configuration, when both rat and human goal-reaching was as its peak.

To our knowledge, this is the first investigation into spatial navigation strategies used by humans and rats by directly comparing them to reinforcement learners on the same task. Previous research has also investigated neural implementations of reinforcement learning strategies in either humans (Gläscher et al., 2010; Simon and Daw, 2011) or rats (Miller et al., 2017), but typically this has focussed on the dichotomy of model-free and model-based methods. The outcome of these studies generally point to the existence of both model-free and model-based strategies in the brain (Daw et al., 2005, 2011). However, these studies have not used an intermediary approach such as the successor representation, which has been shown to provide a link between model-based and model-free mechanisms (Russek et al., 2017).

Our findings are in agreement with previous work comparing the decision making of humans to successor representation agents (Momennejad et al., 2017), and supports the hypothesis that the brain utilises a predictive map of its environment to facilitate spatial decision making (Stachenfeld et al., 2017). Whilst this theory provides a general framework that can be applied to various problem spaces, it would be reasonable to question how it could actually be implemented in neurons. In particular, the initial discretisation of the state space into states is subjective, and a far from our understanding of how the brain represents physical space. In the next chapter, we will introduce a framework that integrates elements of our understanding of the hippocampal representations into a neurally plausible successor representation model of place and grid cell firing.

Chapter 3

Neurobiological successor features for spatial navigation

3.1 Introduction

It has been proposed that the hippocampus encodes a successor representation of space (Stachenfeld et al., 2017). As we explored in the previous two chapters, this formulation typically involves discretisation of the environment into a grid of locations, within which the successor representation can be learnt by transitioning around the grid of states. This fixed grid-world renders it hard to make predictions about how environmental manipulations, such as dimensional stretches, would immediately affect hippocampal representations. Furthermore, in very large state spaces, estimating the successor representation for every state becomes an increasingly difficult and costly task. Instead, using a set of features to approximate location would allow generalisation across similar states and circumvent this curse of dimensionality (Barreto et al., 2017). Indeed, it is clear from electrophysiological studies of the neural circuits supporting navigation that the brain does not represent space as a grid of discrete states, but rather uses an array of spatially sensitive neurons.

Boundary responsive neurons are found throughout the hippocampal formation, including 'border cells' in superficial medial entorhinal cortex (mEC) (Solstad et al., 2008) and boundary vector cells (boundary vector cells) in subiculum (Hartley et al., 2000; Barry et al., 2006; Lever et al., 2009). Since these neurons effectively provide a representation of the environmental topography surrounding the animal and – in the case of the mEC - are positioned to provide input to the main hippocampal subfields (Zhang et al., 2014), it seems plausible that they might function as an efficient substrate for a successor representation.

Thus the aim of this chapter is to build and evaluate a successor representation (SR) based upon known neurobiological features in the form of boundary vector cells (BVCs) (Hartley et al., 2000; Barry et al., 2006; Solstad et al., 2008; Lever et al., 2009). Not only does this provide an efficient foundation for solving goal-directed spatial navigation problems, we show it provides an explanation for electrophysiological phenomena currently unaccounted for by the standard successor representation model (Stachenfeld et al., 2017).

3.2 Model framework

We generated a population of boundary vector cells following the specification used in previous iterations of the boundary vector cell model (Hartley et al., 2000; Barry and Burgess, 2007; Grieves et al., 2018).

Using a set of 160 boundary vector cells, each position or state s in the environment corresponds to a vector of boundary vector cell firing rates $\boldsymbol{f}(s) = [f_1(s), f_2(s), ..., f_{160}(s)]$ (Figure 3.1). We use a tilde $\tilde{}$ to indicate variables constructed in the boundary vector cell feature space of $\boldsymbol{f}$. All simulations were implemented using the same set of 160 BVCs with parameters $\sigma_{ang} = 11.25$, $\beta = 12$ and $\xi = 8$ (equation 1.3). We used 16 preferred angles: ϕ = [0°, 22.5°, 45° 67.5°, 90°, 112.5°, 135°, 157.5°, 180°, 202.5°, 225°, 247.5°, 270°, 292.5°, 315°,337.5°] at 10 preferred distances d = [3.3cm, 10.2cm, 17.5cm, 25.3cm, 33.7cm, 42.6cm, 52.2cm, 62.4cm, 73.3cm, 85.0cm] chosen to provide uniform overlap between consecutive angular and radial tunings.

By learning a successor representation $\tilde{M}$ among these boundary vector cell features we can use linear function approximation of the value function to learn a set of weights $\tilde{\boldsymbol{R}}(s) = [\tilde{R}_1(s), \tilde{R}_2(s), ..., \tilde{R}_n(s)]$ such that:

$$V(s, \tilde{\boldsymbol{R}}) = \tilde{\boldsymbol{\psi}}(s)^{\intercal}\tilde{\boldsymbol{R}} = \sum_{i=1}^{n} \tilde{\psi}_i \tilde{R}_i \tag{3.1}$$

Where $\intercal$ denotes the transpose and $\tilde{\boldsymbol{\psi}}(s) = \tilde{M}\boldsymbol{f}(s)$ is the vector of successor features constructed using the boundary vector cells as basis features. Analogous to the discrete state-space case where the successor matrix M provides a predictive mapping from the current state to the expected future states, the successor matrix $\tilde{M}$ provides a predictive mapping from current BVC firing rates $\boldsymbol{f}(s)$ to expected future BVC firing rates. Importantly, just as in the discrete case, $\tilde{M}$ and $\tilde{\boldsymbol{R}}$ can be learnt online using temporal difference learning rules:

$$\tilde{M} \leftarrow \tilde{M} + \alpha_{\tilde{M}}[\boldsymbol{f}(s_t) + \gamma\tilde{\boldsymbol{\psi}}(s_{t+1}) - \tilde{\boldsymbol{\psi}}(s_t)]\boldsymbol{f}(s_t)^{\intercal} \tag{3.2}$$

$$\tilde{\boldsymbol{R}} \leftarrow \tilde{\boldsymbol{R}} + \alpha_{\tilde{\boldsymbol{R}}}[R_t + \gamma V(s_{t+1}, \tilde{\boldsymbol{R}}) - V(s_t, \tilde{\boldsymbol{R}})]\tilde{\boldsymbol{\psi}}(s_t) \tag{3.3}$$

where $\alpha_{\tilde{M}}$ and $\alpha_{\tilde{\boldsymbol{R}}}$ are the learning rates for the successor representation $\tilde{M}$ and weight vector $\tilde{\boldsymbol{R}}$ respectively. In particular, equation 3.3 follows from linear function approximation of the value function using semi-gradient descent (Sutton and Barto, 2018) and is derived in appendix A. Since equation 3.2 is independent of reward R_t, the model is still able to capture the structure of the environment in the absence of reward ($\tilde{\boldsymbol{R}} = 0$) by learning the successor matrix $\tilde{M}$. In this manner it inherently describes spatial latent learning as described in rodents (Tolman, 1948).

Consequently, we can learn through experience which boundary vector cells are predictive of others by estimating the successor representation matrix $\tilde{M}$. More precisely, given the agent is at position s with boundary vector cell population firing rate vector $\boldsymbol{f}(s)$, $\tilde{\boldsymbol{\psi}}(s) = \tilde{M}\boldsymbol{f}(s) = \mathbb{E}[\sum_{t=0}^{\infty}\gamma^t \boldsymbol{f}(s_t)|s_0 = s]$ represents the expected sum of future population firing rate vectors, exponentially discounted into the future by the parameter $\gamma \in [0,1]$.

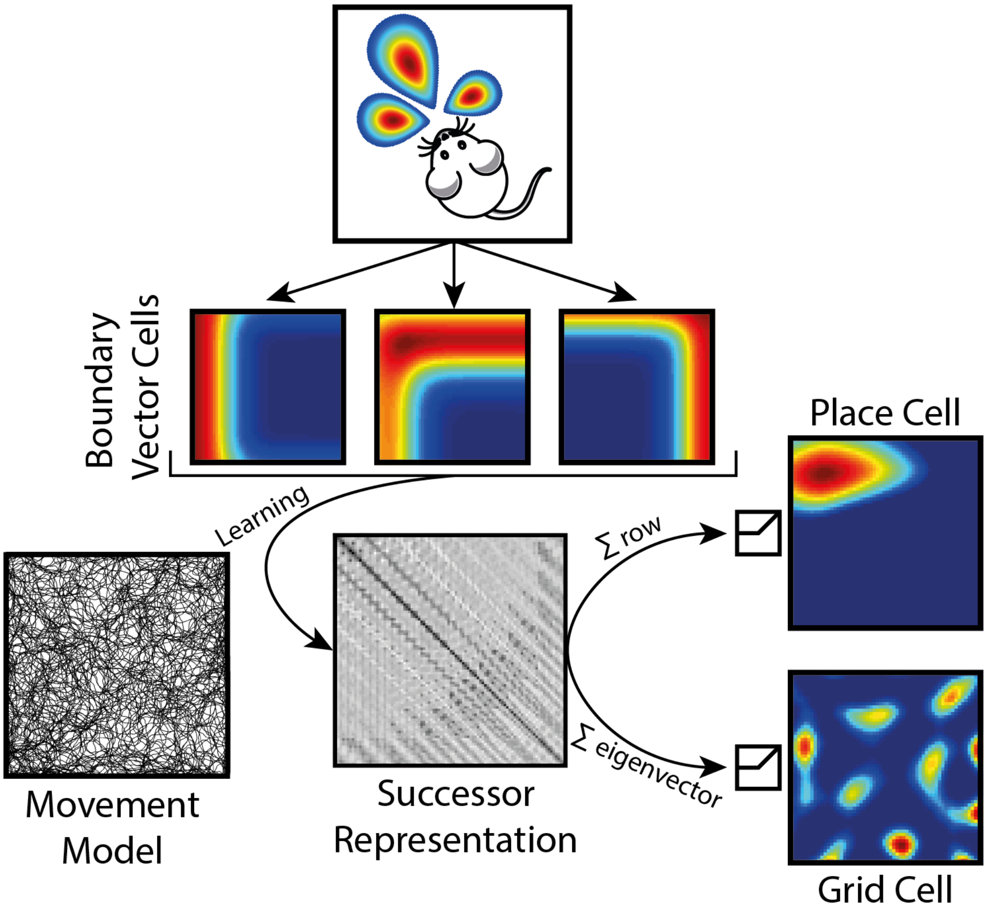


Figure 3.1: Schematic of the BVC-SR model. Boundary vector cells (boundary vector cells), which track the agent's allocentric distance and direction from environmental boundaries, are used as basis features for a successor representation ($\tilde{M}$). The agent's behaviour is generated using a rodent-like movement model with the successor matrix being updated incrementally at each 50Hz time step. Following from previous analyses of the successor matrix - thresholded sums of the boundary vector cell features, weighted by rows of the successor representation matrix, yield unimodal firing fields with characteristics similar to CA1 place cells. Similarly, thresholded eigenvectors of the successor matrix reveal spatially periodic firing patterns similar to medial entorhinal grid cells.

This contrasts with previous implementations of the successor representation where rows and columns of the matrix M correspond to particular states. Here, rows and columns of the successor representation matrix correspond to particular boundary vector cells instead. Specifically, the element $\tilde{M}_{ij}$ can be thought of as a weighting for how much the j^{th} boundary vector cell predicts the firing of the i^{th} boundary vector cell in the near future. Thus, whilst boundary vector cell firing $\boldsymbol{f}$ depends on the environmental boundaries, the successor representation matrix $\tilde{M}$ and consequently successor features $\tilde{\boldsymbol{\psi}}$ are policy dependent meaning they are shaped by behaviour.

In order to generate the behaviour used for learning, we utilised a motion model designed to mimic foraging behaviour of rodents (Raudies and Hasselmo, 2012). The motion model is characterised by a random walk with a preference to follow the boundaries of an environment (figure 3.1). It has previously has been used to investigate grid cell representations in deep reinforcement learning agents (Banino et al., 2018). Trajectories were sampled at a frequency of 50hz and the learning update in equation 3.2 was processed at every time point.

3.2.1 Place cells

Similar to the boundary vector cell model (Hartley et al., 2000), the firing of each simulated place cell F_i in a given location s is proportional to the thresholded, weighted sum of the boundary vector cells connected to it:

$$F_i(s) \propto \Theta(\sum_j \tilde{M}_{ij} f_j(s) - T) \tag{3.4}$$

where T is the cell's threshold and

$$\Theta(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

The weights in the sum (equation 3.4) correspond to a row of the successor representation matrix $\tilde{M}$ and refer to the individual contributions that make a particular

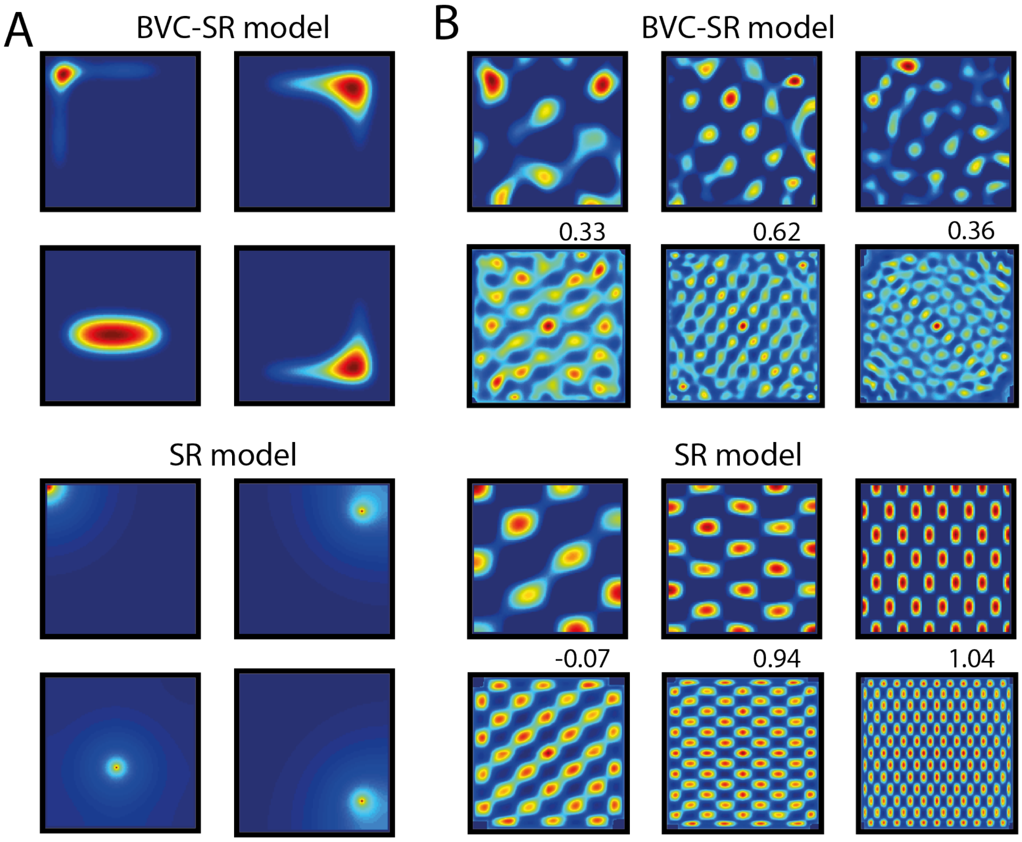


Figure 3.2: Typical place and grid cells generated by the BVC-SR and standard successor representation models. (A) Like rodent CA1 place cells, BVC-SR place cells (top) in the open field are non-uniform, irregular, and often conform to the geometry of the environment. In contrast standard successor representation place cells (bottom) are characterised by smooth, circular fields. (B) Grid cells in both the BVC-SR (1st row) and successor representation models (3rd row) are produced by taking the eigenvectors of the successor representation matrix. The corresponding spatial autocorrelograms (2nd and 4th rows) are used to assess the hexagonal periodicity (gridness) of the firing patterns, shown above each spatial autocorrelogram.

boundary vector cell (encoded by that row) likely to fire in the near future. Thus, assuming homogeneous behaviour, sets of boundary vector cells with overlapping fields will typically exhibit mutually strong positive weights, resulting in the formation of place fields at their intersection (Figure 3.2A). Place cell rate maps were plotted using the weighted sum of their input boundary vector cells' rate maps and setting the threshold T to 80% of the place cell's maximum activation.

3.2.2 Grid cells

Grid cells in the model are generated by taking the eigendecomposition of the successor representation matrix $\tilde{M}$ and thus represent a low-dimensional embedding of the successor representation. Similar to the place cells, the activity of each simulated grid cell G_i is proportional to a thresholded, weighted sum of boundary vector cells. However, for the grid cells, the weights in the sum correspond to particular eigenvector $\tilde{\boldsymbol{v}}^i$ of the successor representation matrix $\tilde{M}$ and the firing is thresholded at zero to only permit positive grid cell firing rates.

$$G_i(s) \propto \Theta(\sum_j \tilde{v}^i_j f_j(s)) \tag{3.6}$$

This gives rise to spatially periodic firing fields such as those observed in Figure 3.2B.

3.3 Simulations

Following Stachenfeld and colleagues (Stachenfeld et al., 2017), we propose that hippocampal place cells could encode the successor features $\tilde{\boldsymbol{\psi}}$ of boundary vector cells to facilitate decision making during spatial navigation. Importantly, due to the disassociation of $\tilde{\boldsymbol{\psi}}$ and reward weights $\tilde{\boldsymbol{R}}$ in the computation of value (equation 3.1), the model facilitates the latent learning phenomena observed by Tolman and Honzik (1930). This is due to the independent learning of $\tilde{\boldsymbol{\psi}}$ irrespective of whether reward is present. The model also provides an efficient platform for solving reward revaluation problems by simply changing the reward weights $\tilde{\boldsymbol{R}}$, .

Like real place cells and those generated by the standard-SR model (Stachenfeld et al., 2017), place cells simulated with the BVC-SR model respect the transition statistics of the environment and thus do not extend through environmental boundaries. However due the nature of the underlying boundary vector cell basis features, the simulated place cells also exhibit characteristics of hippocampal place cells that are unaccounted for by the standard-SR model. For example, in the standard-SR model, place cell firing in a uniformly sampled open field environment tends to be characterised by circular smoothly decaying fields (Stachenfeld et al., 2017). In contrast, BVC-SR derived place fields – like real place cells and those from the boundary vector cell model (Muller et al., 1987; Hartley et al., 2000) - are elongated along environmental boundaries and generally conform to the shape of the enclosing space (Figure 3.2A).

3.3.1 Dimensional stretches

Most importantly, the use of a boundary vector cell basis set provides a means to predict how the model will respond to instantaneous changes in the structure of the environment. In Stachenfeld et al. (2017), the states available to an agent were distinct from the environmental features that constrained the allowed transitions. Thus, insertion of a barrier into an environment had no immediate effect on place or grid fields – changes in firing fields would accumulate through subsequent exploration and learning causing M to be updated. However, biological results indicate that place cell activity is modulated almost immediately by changes made to the geometry of an animal's environment (O'Keefe and Burgess, 1996; Hartley et al., 2000; Lever et al., 2002; Barry et al., 2006; Barry and Burgess, 2007). Since boundary vector cell activity is defined relative to environmental boundaries, manipulations made to the geometry of an environment produce immediate changes in the activity of place cells without any change to the successor representation matrix $\tilde{M}$. Thus, similar to the standard boundary vector cell model, elongation or compression of one or both dimensions of an open field environment distorts place cell firing in a commensurate fashion (Figure 3.3A-C), as has been seen in rodents (O'Keefe and Burgess, 1996). As a result, the basic firing properties of BVC-SR place cells – such as field size – are relatively preserved between manipulations (Figure 3.3D).

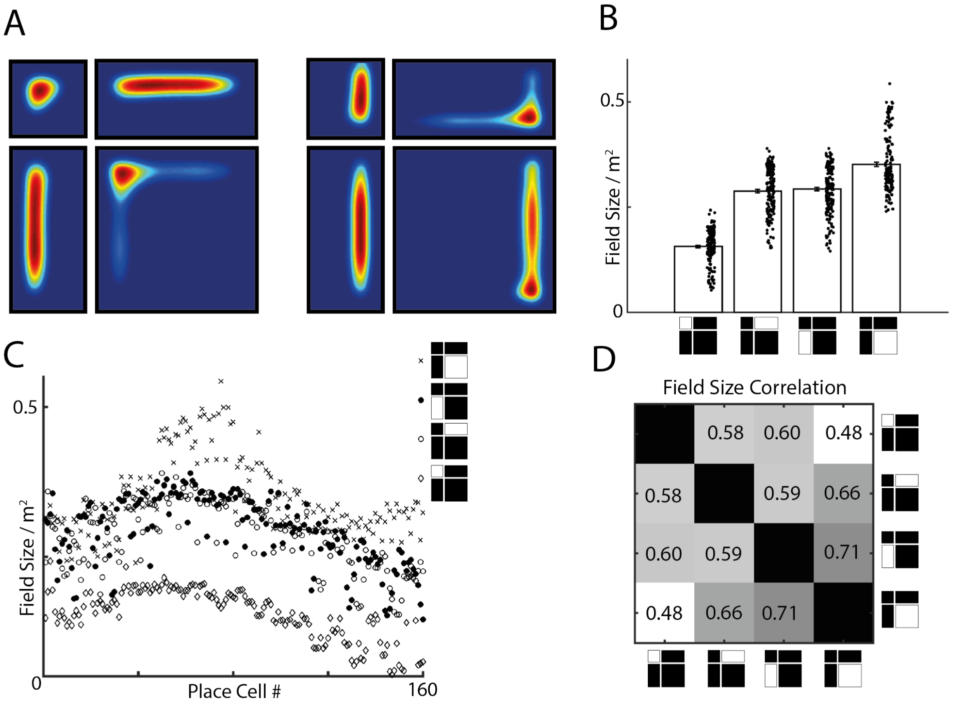


Figure 3.3: BVC-SR derived place cells deform in response to geometric manipulations made to the environment. Scaling one or both axes of an environment produces commensurate changes in the activity of BVC-SR place cells (A). Such that firing field size scales proportionally with environment size (B,C) while the relative size of place fields is largely preserved between environments – Pearson correlation coefficient shown (D).

3.3.2 Barrier insertion

The introduction of internal barriers into an environment provides a succinct test for geometric theories of spatial firing and has been studied in both experimental and theoretical settings. Indeed, the predictable allocentric responses of biological boundary vector cells to inserted barriers provides some of the most compelling evidence for their existence (Lever et al., 2009; Poulter et al., 2018). In CA1 place cells, barrier insertion promotes an almost immediate duplication of place fields (Muller and Kubie, 1987) which may then be then lost or stabilised during subsequent exploration (Barry et al., 2006; Barry and Burgess, 2007). The BVC-SR model provided a good account of empirical data, exhibiting similar dynamic responses. Barrier insertion caused 23% of place cells (32 of 160) to immediately form an additional field, one being present on either side of the barrier (Figure 3.4A). Following further exploration, 19% of these (7 out of 32) gradually lost one of the duplicates – a modification reflecting updates made to $\tilde{M}$ resulting from changes in behaviour due to the barrier (Figure 3.4B) (Barry and Burgess, 2007). Upon removal of the barrier, the simulated place cells reverted more or less to their initial tuning fields prior to barrier insertion, with minor differences due to the updated successor representation $\tilde{M}$.

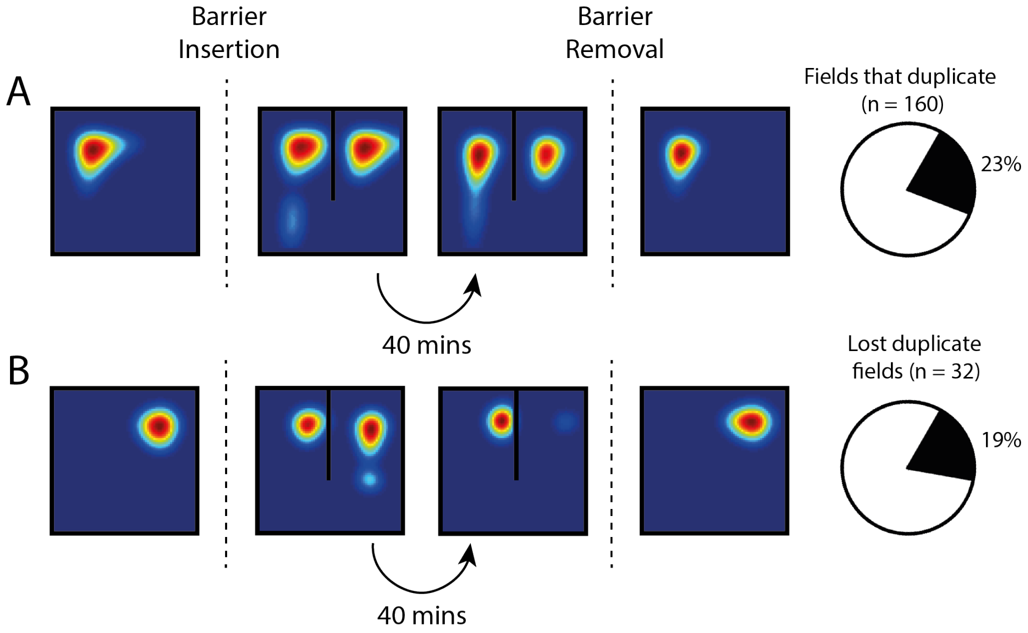


Figure 3.4: Insertion of an additional barrier into an environment can induce duplication of BVC-SR place fields. (A) In 23% of place cells, barrier insertion causes immediate place field duplication. In most cases (81%) the duplicate field persists for the equivalent of 40 minutes of random foraging (learning update occurs at 50Hz) (B) In some cases (19%) one of the duplicate fields – not necessarily the new one – is lost during subsequent exploration. Similar results have been observed in vivo (Barry et al., 2006).

3.3.3 The influence of environmental geometry

Stachenfeld et al. (2017) previously demonstrated that eigendecomposition of the successor matrix M produced spatially periodic firing fields resembling mEC grid cells. Examining the eigenvectors of $\tilde{M}$, from the BVC-SR model, we found that these too resembled the regular firing patterns of grid cells (Figure 3.2B). Indeed, while there was no difference in the hexagonal regularity of BVC-SR and standard-SR eigenvectors (mean gridness ± SD: -0.28 ± 0.35 vs. -0.27 ± 0.60; t(318)=0.14, p=0.886), the eigenvectors from the BVC-SR model exhibit less elliptic grid fields (mean field ellipticity ± SD: 0.59 ± 0.23 vs. 0.75 ± 0.25; t(318)=-5.93; $p < 0.001$; Figure 3.5). Since neither the BVC-SR or standard-SR models yield exclusively hexagonal grid patterns, ellipticity of the eigenvector rate maps was calculated by thresholding the spatial autocorrelogram at a value of 0.2 and identifying the central peak. Next, the eccentricity e of this central peak was used as a measure of the grid ellipticity:

$$e = \sqrt{1 - \frac{a^2}{b^2}} \tag{3.7}$$

where a and b are the lengths of the longer and shorter axis of the central peak.

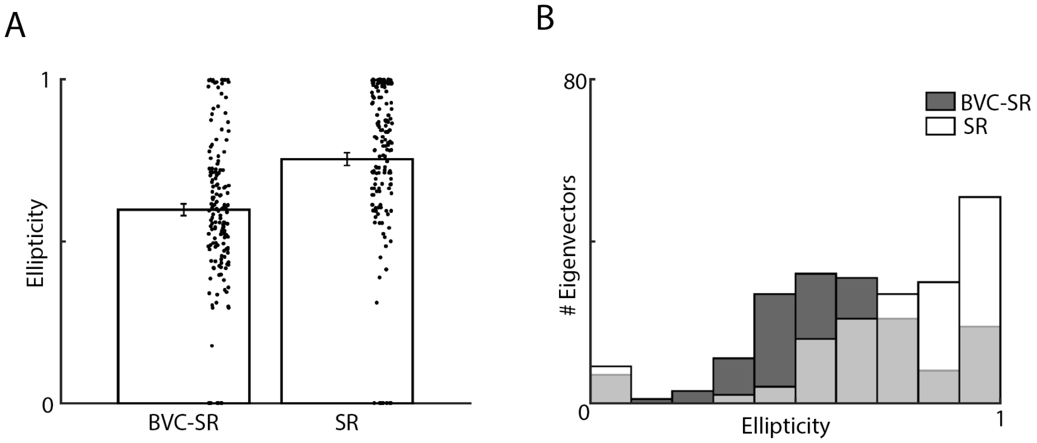


Figure 3.5: Grid fields generated using eigenvectors from the BVC-SR model are less elliptic than those from the standard-SR model. Lower values indicate more circular fields and larger values indicate more elliptic fields, with a value of 0 indicating a perfect circle. (A) Grid fields generated using the BVC-SR model had significantly lower ellipticity than the standard-SR model (mean field ellipticity ± SD: 0.59 ± 0.23 vs. 0.75 ± 0.25; t(318)=-5.93; $p < 0.001$), and were similar to observations of real grid cells (Krupic et al., 2015). (B) Histogram of the grid field ellipticity ($N = 160$ eigenvectors)

The grid patterns generated using the BVC-SR approach also had a larger variability in field firing rates than the standard-SR method (Stachenfeld et al., 2017). Firing rate variability of the eigenvector rate maps was analysed following the method of (Ismakov et al., 2017). Grid fields were identified using the watershed transform of each eigenvector rate map, and the coefficient of variability (CV) was calculated as the standard deviation of these peaks, divided by the mean of the peaks. The observed CVs of eigenvectors from the BVC-SR model (BVC-SR vs SR, mean CV

± SD: 0.48 ± 0.11 vs 0.14 ± 0.11; t(318)=26.5; $p < 0.001$; Figure 3.6), was similar to that observed in real grid cells (Ismakov et al., 2017).

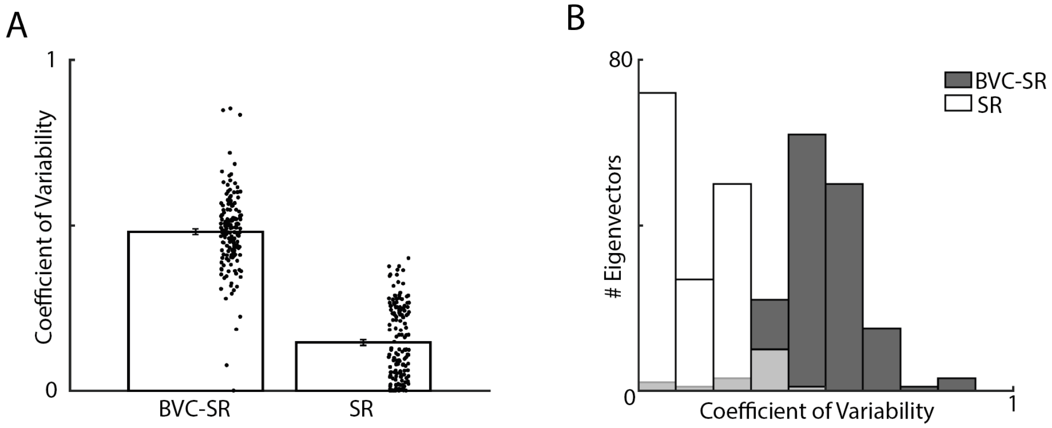


Figure 3.6: Grid fields generated using eigenvectors from the BVC-SR model are more exhibit more firing rate variability than the standard-SR model. Following the method of Ismakov et al., (2017), the peak firing rates of grid fields was used to compute a coefficient of variability for each eigenvector (CV; SD divided by mean). (A) The CV for eigenvectors produced by the BVC-SR model were significantly larger than that observed in the standard-SR model (mean CV ± SD: 0.48 ± 0.11 vs 0.14 ± 0.11; t(318)=26.5; $p < 0.001$), and similar to that observed in real grid cells (Ismakov et al., 2017). (B) Histogram of the CV for each of the models ($N = 160$ eigenvectors).

Empirical work has shown that grid-patterns are modulated by environmental geometry, the regular spatial activity becoming distorted in strongly polarised environments (Derdikman et al., 2009; Krupic et al., 2015; Stensola et al., 2015). Grid-patterns derived from the standard-SR eigenvectors also exhibit distortions comparable to those seen experimentally. Thus, we next examined the regularity of BVC-SR eigenvectors derived from successor representation matrices trained in square and trapezoid environments. As with rodent data (Krupic et al., 2015) and the standard-SR model (Stachenfeld et al., 2017), we found that grid-patterns in the two halves of the square environment were considerably more regular than those derived from the trapezoid (mean correlation between spatial autocorrelograms ± SD: 0.68 ± 0.18 vs. 0.47 ± 0.15, $t(317) = 10.99$, $p < 0.001$; Figure 3.7B). Furthermore, BVC-SR eigenvectors that exceeded a shuffled gridness threshold (Barry and Burgess, 2017) – and hence were classified as grid cells – were more regular in the square than the trapezoid (mean gridness ± SD: 0.37 ± 0.17 vs. 0.10 ± 0.09;

$t(24) = 4.87$, $p < 0.001$; Figure 3.7C). In particular, as had previously been noted in rodents (Krupic et al., 2015), the regularity of these ‘grid cells’ was markedly reduced in the narrow end of the trapezoid compared to the broad end (mean gridness ± SD: -0.30 ± 0.19 vs. 0.16 ±0.23; $t(22) = -5.45$, $p < 0.001$; Figure 3.7D), a difference that did not exist in the two halves of the square environment (mean gridness ± SD: 0.19 ± 0.25 vs. 0.22 ± 0.36; $t(26) = -0.28$, $p = 0.78$).

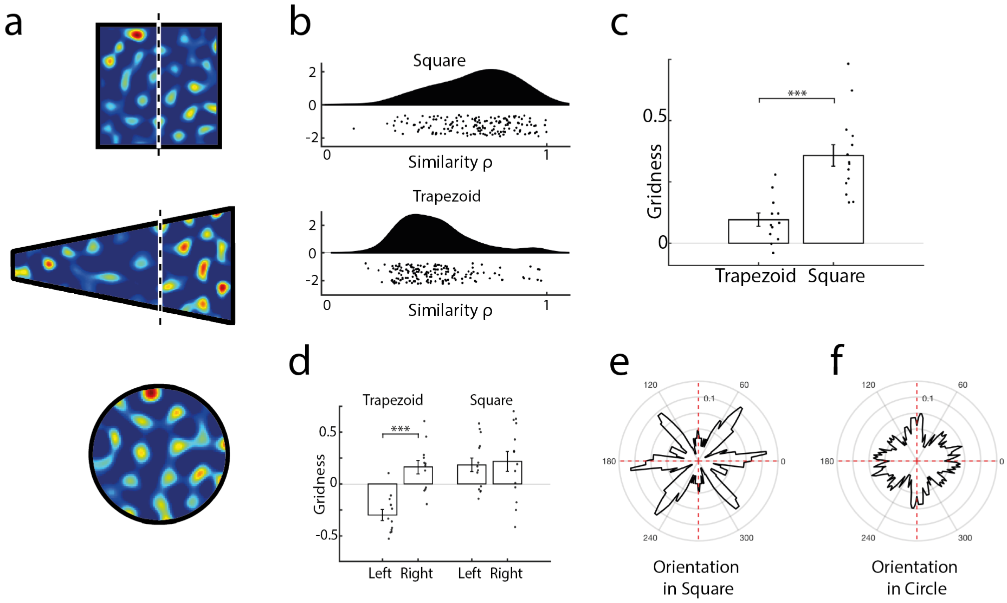


Figure 3.7: BVC-SR grid-patterns are influenced by environmental geometry. (A) Eigenvectors of the boundary vector cell- successor representation can be used to model grid cells firing patterns in a variety of different shaped enclosures (white line indicates division of square and trapezoid into halves of equal area). (B) Grid-patterns are more similar in the two halves of the square environment than in the two halves of the trapezoid (mean Pearson’s correlation between spatial autocorrelograms ± SD: 0.68 ± 0.18 vs. 0.47 ± 0.15, $t(317) = 10.99$, $p < 0.001$), similar results have been noted in rodents (Krupic et al., 2015). (C) ‘Grid cells’ (grid- patterns that exceed a shuffled gridness criteria) are more hexagonal in the square environment than the trapezoid (mean gridness ± SD: 0.37 ± 0.17 vs. 0.10 ± 0.09; $t(24) = 4.87$, $p < 0.001$), (D) the narrow half of the trapezoid being less regular than the wider end (mean gridness ± SD: -0.30 ± 0.19 vs. 0.16 ±0.23; $t(22) = -5.45$, $p < 0.001$). The axes of ‘grid cells’ are more polarised (less uniform) in a square (E) than circular environment (F) (D_{KL}(Square$||$Uniform) = 0.17, D_{KL}(Square$||$Uniform) = 0.04; Bayes Factor = 1.00×10^{-6}).

Rodent grid-patterns have been shown to orient relative to straight environmental boundaries – tending to align to the walls of square but not circular environments (Krupic et al., 2015; Stensola et al., 2015). In a similar vein, we saw that firing patterns of simulated grid cells also were more polarised in a square than a circular environment, tending to cluster around specific orientations (Figure 3.7EF). To illustrate this, we used the Kullback-Leibler divergence (D_{KL}) to measure the difference between the distribution of grid orientations and a uniform distribution. The Kullback-Leibler divergence is a measure of how different a probability distribution P (i.e. grid orientations) is from a reference distribution Q (i.e. uniform distribution) defined on the same probability space, and is calculated as:

$$D_{KL} = \sum_{x} P(x) \frac{P(x)}{Q(x)} \tag{3.8}$$

We found the grid orientations in the circular environment were much closer to uniform (D_{KL}(Circle$||$Uniform) $= 0.04$ vs. D_{KL}(Square$||$Uniform) $= 0.17$), and significantly better explained by an underlying uniform distribution as opposed to the grid orientations in the square environment (Bayes Factor Analysis: Bayes Factor = 1.00×10^{-6}; Kass and Raftery (1995)).

Finally, the activity of grid cells recorded whilst a rodent explores a compartmentalised maze (Figure 3.8A) have been shown to fragment into repeated submaps for similar compartments traversed in the same direction (Derdikman et al., 2009). We simulated the BVC-SR model in a similar maze (Figures B-C) and found that the eigenvector patterns also fragment into repeated submaps for alternating internal arms of the maze (Figure 3.8D). Consequently, the Pearson's correlation matrix between eigenvector patterns on different arms of the maze exhibits a strong checkboard-like appearance (Figure 3.7E), exemplifying the repetition of alternated submaps in a manner more similar to the rodent data (Derdikman et al., 2009) than the standard-SR model (Stachenfeld et al., 2017).

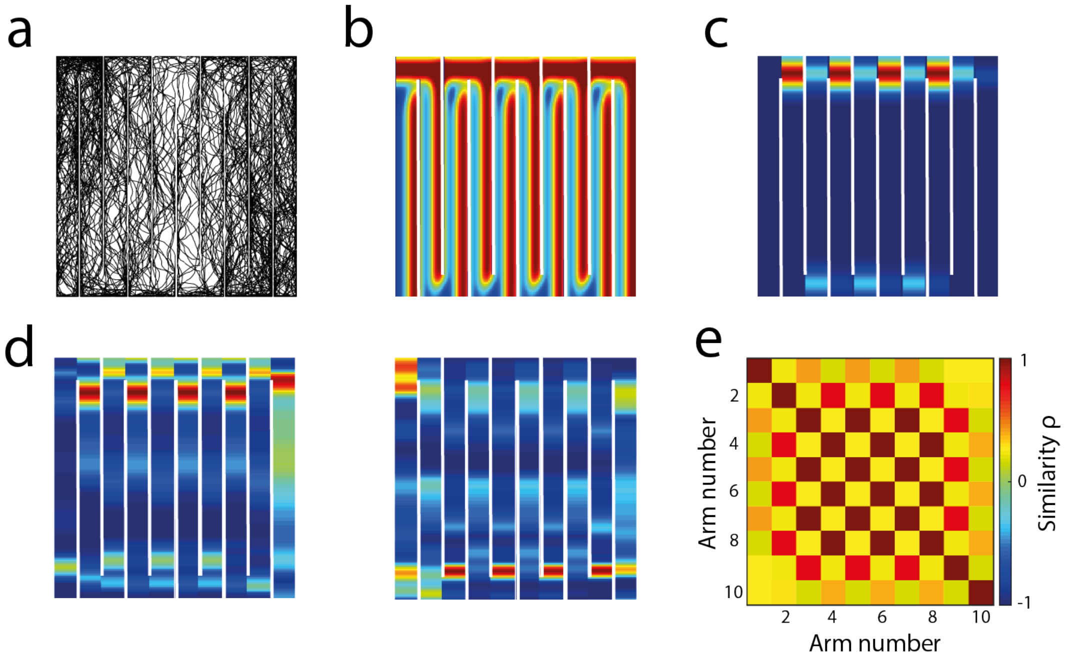


Figure 3.8: (A) The trajectory used to facilitate the learning of the successor representation between boundary vector cells (BVCs) in a compartmentalised maze. (B) An example BVC firing field tuned to boundaries directed towards the top-right corner of the maze. (C) An example BVC-SR place cell displaying repeated firing fields across alternate arms of the maze. (D) The BVC-SR eigenvector grid patterns are fragmented in the compartmentalised maze and repeat across alternating maze arms as has been observed in rodents (Derdikman et al., 2009). (E) The Pearson's correlation matrix between the grid patterns on different arms of the maze has a checkerboard-like appearance due to the strong similarity between alternating internal channels of the maze ($n = 160$ eigenvectors).

3.4 Discussion

The model presented here links the boundary vector cell model of place cell firing with a successor representation to provide an efficient platform for using reinforcement learning to navigate space. The work builds upon previous implementations of the successor representation by using a basis set of known neurobiological features - boundary vector cells, which have been observed in the hippocampal formation (Barry et al., 2006; Solstad et al., 2008; Lever et al., 2009) and can be derived from optic flow (Raudies and Hasselmo, 2012). As a consequence, the place cells generated using the BVC-SR approach presented here produce more realistic fields that conform to the shape of the environment and respond immediately to environmental manipulations. Comparable to previous successor representation implementations, the eigenvectors of the successor representation matrix $\tilde{M}$ display grid cell like periodicity when projected back onto the boundary vector cell state space, with reduced periodicity in polarised enclosures such as trapezoids. Further, likely due to the experiential learning and the natural smoothness of the BVC basis features, the eigenvectors from the BVC-SR model exhibit more realistic variations among grid fields, resulting in a model of grid cells that is more similar to biological recordings than previous implementations of the SR. This form of eigendecomposition is similar to other dimensionality reduction techniques that have been used to generate grid cells from populations of idealised place cells with a generalised Hebbian learning rule (Oja, 1982; Dordek et al., 2016). Previously, low dimensional encodings such as these have been shown to accelerate learning and facilitate vector-based navigation (Gustafson and Daw, 2011; Banino et al., 2018).

The model extends upon the boundary vector cell model of place cell firing (Hartley et al., 2000; Barry et al., 2006; Barry and Burgess, 2007) by also providing a means of predicting how environmental boundaries might affect the firing of grid cells. Furthermore, whilst both models produce similar place cells if the agent samples the environment uniformly, the policy dependence of the BVC-SR model provides a mechanism for estimating how behavioural biases will influence place cell firing. These models both use boundary vector cells as the basis for allocentric place representations in the brain. However, theoretical (Byrne et al., 2007; Bicanski and Burgess, 2018) and recent experimental evidence (Hinman et al., 2019; Gofman et al., 2019) suggests that egocentric boundary cells may provide a link between the egocentric perception and allocentric representation of space.

The focus of this model has centred on the representation of successor features in the hippocampus during the absence of environmental reward. However a key feature of successor representation models is their ability to adapt flexibly and efficiently to changes in the reward structure of the environment (Dayan, 1993; Russek et al., 2017; Stachenfeld et al., 2017). This is permitted by the independent updating of reward weights (equation 3.3) combined with its immediate effect on the computation of value (equation 3.1). Reward signals analogous to that used in the model have been shown to exist in the orbitofrontal cortex of rodents (Sul et al., 2010), humans (Gottfried et al., 2003; Kringelbach, 2005) and non-human primates (Tremblay and Schultz, 1999). Meanwhile a candidate area for integrating orbitofrontal reward representations with hippocampal successor features to compute value could be anterior cingulate cortex (Shenhav et al., 2013; Kolling et al., 2016). Finally, the model relies on a prediction error signal for learning both the reward weights and successor features (equations 3.2 and 3.3). Whilst midbrain dopamine neurons have long been considered a source for such a reward prediction error (Schultz et al., 1997), mounting evidence suggests they may also provide the sensory prediction error signal necessary for computing successor features with temporal-difference learning (Chang et al., 2017; Gardner et al., 2018).

Successor features have been used to accelerate learning in tasks where transfer of knowledge is useful, such as virtual and real world navigation tasks (Barreto et al., 2017; Zhang et al., 2017). Whilst the successor features used in this paper were built upon known neurobiological spatial neurons, the framework itself could be applied to any basis of sensory neurons that are predictive of reward in a task. Thus, the framework could be adapted to use basis features that are receptive to the frequency of auditory cues (Aronov et al., 2017), or even the size and shape of birds (Constantinescu et al., 2016).

In summary, the model describes the formation of place and grid fields in terms the geometric properties and transition statistics of the environment, whilst providing an efficient platform for goal-directed spatial navigation. This has particular relevance for the neural underpinnings of spatial navigation, although the framework itself could be applied to other basis sets of sensory features.

Chapter 4

General Discussion

4.1 Humans, rats and predictive maps

Using a combination of likelihood and behavioural similarity analyses we showed that the spatial behaviour of humans and rats was more indicative of a successor representation mechanism than either model-based or model-free. The behavioural similarity was robust enough to build a reinforcement learning agent classifier, which predominantly predicted biological behaviour to a successor representation agent.

One scenario that was unexplored in this work is the possibility of a hybrid agent that combines the predictions of a model-free learning system and a more structure oriented approach such as a successor representation or model-based system. Such a hybrid agent could then arbitrate between the action predictions of the dual systems based on their recent reliability (Daw et al., 2005), in order to benefit from both the efficiency of the model-free as well as the flexibility of a successor representation/model-based system. Such a hybrid architecture has been hypothesised to map onto the striatum/hippocampus respectively (Chersi and Burgess, 2015), and is able to explain a range of behaviours in rodents (Packard and McGaugh, 1996; Pearce et al., 1998; Killcross and Coutureau, 2003) when applied to the relevant state-spaces (Geerts et al., prep). However, highly non-stationary tasks such as the one presented in this thesis would be expected to heavily favour the flexibility provided by the faster learning successor representation or model-based system.

Instead of pooling the maximum likelihood parameter estimates across the group of animals or participants, an alternative possibility would be to fit model parameters to each individual. Similar methods have been successfully used to link model-free and model-based prediction errors to BOLD signal activity in the ventral striatum (Daw et al., 2011). Such an approach would specify to what extent each individual's behaviour is consistent with each model, and might be able to give an indication into the viability of a dual-systems approach by monitoring how the model likelihoods vary within animal/participant across the course the experiment. It is worth noting that the pooling of model parameter estimates across individuals is an important part of generating trajectories for the behavioural similarity analyses, however the comparison of individual fits will be explored further in future work.

Comparing the similarity of trajectories that can vary in length is a non-trivial problem that hinges upon the definition of 'similar'. One might describe two trajectories as similar because at some point they traverse the same section of space. Conversely, two trajectories might be identical but slightly offset from each other so that they never overlap. Our definition of similar relied on comparing trajectories that start at the same position, and used measures designed to capture both translational invariance (i.e. rotational velocity and diffusivity) and larger scale complexity (i.e. tortuosity). In particular, rotational velocity and diffusivity were calculated along a trajectory and could be useful at providing real-time descriptions of spatial behaviour.

Future research could put more emphasis on improving the generality of the models used by model-based methods. For example, the model-based algorithm we implemented had privileged knowledge about the location of reward and connectivity of the states before it even entered the environment. One of the elegant features of a successor representation is how little it assumes about the environment - the transition structure, in the form of the successor representation, is acquired from experience. Indeed, it may be useful to design model-based planning procedures that are implemented on the successor representation itself. To facilitate this, the successor matrix could be normalised to form a transition matrix that captures more of the temporal dynamics than just one-step transitions.

It is important to note that all of the reinforcement learning agent behaviour was unable to recapitulate the goal-reaching performance of the biological data. In terms of 'what is lacking' in these algorithms that a brain might implement whilst solving this task, one issue resides in the exploration dynamics used by the agents. In order to detect behavioural differences between different algorithms, it is important to force relearning by perturbing either the transition or reward structure of the environment. However in response to these perturbations, the agent - whether it is implementing the ε-greedy or softmax policy - is only equipped with a weak exploration strategy driven by chance. In reality, it seems likely that biological explorative behaviour is more focussed in the way it resolves uncertainty in the environment. This would allow it to explore more wisely, and thus exploit future reward faster. In order to implement this *in silico*, the agents would need to incorporate a distribution over the environmental information that they are caching. They could then use the uncertainty in these distributions, along with the point estimates that they currently use, in order to tackle the exploration-exploitation tradeoff in a more informed manner.

4.2 BVC-SR model

Based on the similarity between the biological behaviour and the successor representation, we set out to theorise how this algorithm could actually be implemented in neurons. Under the hypothesis that the hippocampal place cells represent a predictive map of states in the environment, its important to clarify what the actual 'states' are in the brain. Building from experimental work highlighting the strength of boundary inputs to place cells (O'Keefe and Burgess, 1996; Hartley et al., 2000; Barry et al., 2006), we used a set of boundary vector cells (BVCs) as the basis features to a neurobiological successor representation.

While boundary sensitive neurons are found in the entorhinal cortex (Solstad et al., 2008) - a major input to the hippocampus - the relative sparsity of cells with long-range distal tuning has often has often questioned the plausibility of models that use such representations to explain the firing of CA1 place cells (although long-range boundary tunings do exist in the entorhinal cortex - see Koenig et al. (2011) supporting material). Meanwhile, long-range BVCs appear to be more prominent in the subiculum (Lever et al., 2009) - which has long been viewed as a major output of hippocampus. However recent anatomical techniques have found evidence that the subiculum projects to both excitatory and inhibitory CA1 neurons (Sun et al., 2014), and that the strength of this projection is comparable to the direct projection from medial entorhinal cortex to CA1 (Sun et al., 2018). Furthermore, inactivation of the CA1 projecting subicular neurons disrupts the spatial firing of place cells and produces deficits in object-place learning (Sun et al., 2019).

The learning update used in the model (equation 3.2) bears resemblance to Hebbian learning - cells that fire together will have mutually strong weights in the successor matrix. It would be interesting to investigate what effect weight penalties, similar to ridge regression or a BCM learning rule (Bienenstock et al., 1982), would have on the successor matrix. Not only would this represent some form of metabolic regularisation, but it may also lead to the emergence of microcircuits that could more succinctly be described by multiple smaller successor matrices. These matrices could even vary in discount parameters γ to provide value estimates across different time scales, which could fit with an observed gradient in predictive horizons along the hippocampal long-axis in humans (Brunec and Momennejad, 2019).

From a practicality standpoint, it would be useful to be able to compute the successor matrix under certain behavioural assumptions such as uniform sampling of the environment, but without the simulation of specific trajectories. One possibility for this could be to use the overlap between BVC firing fields to estimate a transition matrix in the BVC space, and then use this to approximate the successor representation according to equation 1.13. Similarly, one could compute the covariance between BVC fields. In this case the method of computing grid cells becomes almost equivalent to that of Dordek et al. (2016) but using BVC features instead of place cell features - calculating the eigenvectors of a covariance matrix is one form of principle component analysis (PCA), although variants of PCA that enforce a non-negativity constraint have been found to produce the most hexagonal grid patterns (Sorscher et al., 2019). However, this method of modelling grid cells by computing dimensionality reduction techniques (such as PCA) on a set place cells or BVCs does not represent the whole picture in the entorhinal cortex. Along with grid cells, neurons in the medial entorhinal cortex have been shown to also represent head-direction (Sargolini et al., 2006) and speed (Kropff et al., 2015), as well as conjunctive representation of all three (Hardcastle et al., 2017). Since so far the models of place cells to which the dimensionality reduction has been applied do not take these variables into account, no such representations of direction or speed exist in the modelled grid cells. However it is known that place cell firing does in fact represent the speed of the animal (Huxter et al., 2003), and in linearised environments head-direction also (Mcnaughton et al., 1983), thus a logical extension of the model would be to analyse the effect of including a wider range of spatially informative features, such as head-direction cells (Taube et al., 1990) and the vestibular system. In particular, the vestibular input to grid cells appears to play an important role in maintaining the hexagonal periodicity of the grid patterns, and this signal may be channelled via the medial septum. Inactivation of the medial septum produces a reduction in movement-related theta power and a commensurate reduction in the gridness of grid cells, whilst leaving directional selectivity in the entorhinal cortex intact (Brandon et al., 2011). Meanwhile the spatial firing patterns of hippocampal place cells and entorhinal border cells appear relatively unimpaired by the septal inactivation (Koenig et al., 2011).

Finally, recent work has highlighted the potential of using the communicability between discrete nodes in a network (Estrada and Hatano, 2007) as a metric for guiding intuitive planning in a grid world (Baram et al., 2018). This communicability measure can be computed using the eigendecomposition of the successor representation, and thus confers a functional purpose for grid cells in successor representation models. It remains to be seen whether this method can be extrapolated for use in non-discrete feature spaces such the BVCs used here, but if so it could provide a compelling theoretical insight into how planning could be instantiated in the brain. Indeed the distance and direction signals provided by BVCs may represent an easily actionable feature space in which to readout the outcome of such planning.

4.3 Conclusion

In general, the aim of this thesis was to investigate the existence of predictive maps in rats and humans by examining their spatial behaviour. It went on to theorise how such a representation could actually be implemented in the firing rates of hippocampal neurons. Modelling across these different scales will be crucial if we are to build functional reinforcement learning models of the brain and behaviour that can relate usefully back to neuroscience.

Appendix A

Derivation of equation 3.3

Here we derive the reward weight update in equation 3.3 used for linear function approximation of the value function whilst using the successor features $\tilde{\boldsymbol{\psi}}$ to predict reward. A similar derivation can be found in Sutton and Barto (2018).

To outline the problem: we have the feature vector $\tilde{\boldsymbol{\psi}}$ which is predictive of reward, and we wish to find reward weights $\tilde{\boldsymbol{R}}$ such that $\tilde{V}(s_t,\tilde{\boldsymbol{R}}) = \tilde{\boldsymbol{\psi}}(s_t)^\intercal\tilde{\boldsymbol{R}}$ is an approximation to the true value function $V(s_t)$. Thus we wish to find $\tilde{\boldsymbol{R}}$ that minimises $[V(s_t) - \tilde{V}(s_t,\tilde{\boldsymbol{R}})]^2$ which can be achieved via gradient descent.

$$\begin{aligned}\frac{\partial}{\partial\tilde{\boldsymbol{R}}}[V(s_t) - \tilde{V}(s_t,\tilde{\boldsymbol{R}})]^2 &= \frac{\partial}{\partial\tilde{\boldsymbol{R}}}[V(s_t) - \tilde{\boldsymbol{\psi}}(s_t)^\intercal\tilde{\boldsymbol{R}}]^2 \\ &= 2[V(s_t) - \tilde{\boldsymbol{\psi}}(s_t)^\intercal\tilde{\boldsymbol{R}}] \times \frac{\partial}{\partial\tilde{\boldsymbol{R}}}[\tilde{\boldsymbol{\psi}}(s_t)^\intercal\tilde{\boldsymbol{R}}] \\ &= -2\tilde{\boldsymbol{\psi}}(s_t)[V(s_t) - \tilde{\boldsymbol{\psi}}(s_t)^\intercal\tilde{\boldsymbol{R}}]\end{aligned}$$

This gives us the direction of the positive gradient, thus to move down the slope we must do so proportional to $\tilde{\boldsymbol{\psi}}(s_t)[V(s_t) - \tilde{\boldsymbol{\psi}}(s_t)^\intercal\tilde{\boldsymbol{R}}] = \tilde{\boldsymbol{\psi}}(s_t)[V(s_t) - \tilde{V}(s_t,\tilde{\boldsymbol{R}})]$. As usual, the true value function $V(s_t)$ is unknown, however using the Bellman equation we can use the latest reward signal R_t along with the recursion of expected rewards to bootstrap a best guess $V(s_t) \approx R_t + \gamma\tilde{V}(s_{t+1},\tilde{\boldsymbol{R}})$.

Thus arriving at the update equation 3.3

$$\tilde{\boldsymbol{R}} \leftarrow \tilde{\boldsymbol{R}} + \alpha \tilde{\boldsymbol{\psi}}(s_t)[R_t + \gamma \tilde{V}(s_{t+1}, \tilde{\boldsymbol{R}}) - \tilde{V}(s_t, \tilde{\boldsymbol{R}})]$$

for learning rate α and discount factor γ. For ease of notation and consistency with other chapters, $\tilde{V}(s_t, \tilde{\boldsymbol{R}})$ is referred to as $V(s_t, \tilde{\boldsymbol{R}})$ in the main text.

Bibliography

Andersen, P., Bliss, T. V., and Skrede, K. K. (1971). Lamellar organization of hippocampal excitatory pathways. *Experimental Brain Research*, 13(2):222–238.

Andersen, P., Morris, R., Amaral, D., Bliss, T., and O'Keefe, J. (2006). *The hippocampus book*. Oxford university press.

Aronov, D., Nevers, R., and Tank, D. W. (2017). Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature*, 543(7647):719–722.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M. J., Degris, T., Modayil, J., Wayne, G., Soyer, H., Viola, F., Zhang, B., Goroshin, R., Rabinowitz, N., Pascanu, R., Beattie, C., Petersen, S., Sadik, A., Gaffney, S., King, H., Kavukcuoglu, K., Hassabis, D., Hadsell, R., and Kumaran, D. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433.

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

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., Van Hasselt, H., and Silver, D. (2017). Successor features for transfer in reinforcement learning. *Advances in Neural Information Processing Systems*, 2017-Decem:4056–4066.

Barry, C. and Burgess, N. (2007). Learning in a Geometric Model of Place Cell Firing. *Hippocampus*, 17:786–800.

Barry, C. and Burgess, N. (2017). To be a Grid Cell: Shuffling procedures for determining "Gridness". *bioRxiv*, page 230250.

Barry, C., Hayman, R., Burgess, N., and Jeffery, K. J. (2007). Experience-dependent rescaling of entorhinal grids. *Nature Neuroscience*, 10(6):682–684.

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. *Reviews in the Neurosciences*, 17(1-2).

Barry, C. J. (2007). *Terra Cognita : Representations of Space in the Rodent Hippocampus and Entorhinal Cortex*. PhD thesis.

Bellmund, J. L. S., de Cothi, W., Ruiter, T. A., Nau, M., Barry, C., and Doeller, C. F. (2019). Deforming the metric of cognitive maps distorts memory. *Nature Human Behaviour*, pages 1–12.

Bicanski, A. and Burgess, N. (2018). A neural-level model of spatial memory and imagery. *eLife*, 7(7052):1–3.

Bienenstock, E. L., Cooper, L. N., and Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2(1):32–48.

Boccara, C. N., Sargolini, F., Thoresen, V. H., Solstad, T., Witter, M. P., Moser, E. I., and Moser, M. B. (2010). Grid cells in pre-and parasubiculum. *Nature Neuroscience*, 13(8):987–994.

Brandon, M. P., Bogaard, A. R., Libby, C. P., Connerney, M. A., Gupta, K., and Hasselmo, M. E. (2011). Reduction of theta rhythm dissociates grid cell spatial periodicity from directional tuning. *Science*, 332(6029):595–599.

Brun, V. H., Solstad, T., Kjelstrup, K. B., Fyhn, M., Witter, M. P., Moser, E. I., and Moser, M. B. (2008). Progressive increase in grid scale from dorsal to ventral medial entorhinal cortex. *Hippocampus*, 18(12):1200–1212.

Brunec, I. K. and Momennejad, I. (2019). Predictive Representations in Hippocampal and Prefrontal Hierarchies. *bioRxiv*, page 786434.

Byrne, P., Becker, S., and Burgess, N. (2007). Remembering the past and imagining the future: A neural model of spatial memory and imagery. *Psychological Review*, 114(2):340–375.

Chang, C. Y., Gardner, M., Di Tillio, M. G., and Schoenbaum, G. (2017). Optogenetic Blockade of Dopamine Transients Prevents Learning Induced by Changes in Reward Features. *Current Biology*, 27(22):3480–3486.e3.

Chen, G., Lu, Y., King, J. A., Cacucci, F., and Burgess, N. (2019). Differential influences of environment and self-motion on place and grid cell firing. *Nature Communications*, 10(1).

Chersi, F. and Burgess, N. (2015). The Cognitive Architecture of Spatial Navigation: Hippocampal and Striatal Contributions. *Neuron*, 88:64–77.

Clemens, L. E., Jansson, E. K. H., Portal, E., Riess, O., and Nguyen, H. P. (2014). A behavioral comparison of the common laboratory rat strains Lister Hooded, Lewis, Fischer 344 and Wistar in an automated homecage system. *Genes, Brain and Behavior*, 13(3):305–321.

Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215.

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

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624.

Del Moral, P. (1997). Nonlinear filtering: Interacting particle resolution. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 325(6):653–658.

Derdikman, D., Whitlock, J. R., Tsao, A., Fyhn, M., Hafting, T., Moser, M. B., and Moser, E. I. (2009). Fragmentation of grid cell maps in a multicompartment environment. *Nature Neuroscience*, 12(10):1325–1332.

Doeller, C. F., Barry, C., and Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, 463.

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(MARCH2016):1–36.

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(6954):184–188.

Estrada, E. and Hatano, N. (2007). Communicability in complex networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 77(3).

Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47.

Finlay, B. L. and Darlington, R. B. (1995). Linked regularities in the development and evolution of mammalian brains. *Science*, 268(5217):1578–1584.

Gardner, M. P., Schoenbaum, G., and Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B: Biological Sciences*, 285(1891).

Geerts, J. P., Chersi, F., Stachenfeld, K. L., and Burgess, N. (in prep). Reliability based arbitration of hippocampal and dorsal striatal decision making.

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

Gofman, X., Tocker, G., Weiss, S., Boccara, C. N., Lu, L., Moser, M. B., Moser, E. I., Morris, G., and Derdikman, D. (2019). Dissociation between Postrhinal Cortex and Downstream Parahippocampal Regions in the Representation of Egocentric Boundaries. *Current Biology*, 29(16):2751–2757.e4.

Gottfried, J. A., O'Doherty, J., and Dolan, R. J. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science*, 301(5636):1104–1107.

Grieves, R. M., Duvelle, É., and Dudchenko, P. A. (2018). A boundary vector cell model of place field repetition. *Spatial Cognition and Computation*, 18(3):217–256.

Gustafson, N. J. and Daw, N. D. (2011). Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. *PLoS Computational Biology*, 7(10).

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(7052):801–806.

Hardcastle, K., Maheswaranathan, N., Ganguli, S., and Giocomo, L. M. (2017). A Multiplexed, Heterogeneous, and Adaptive Code for Navigation in Medial Entorhinal Cortex. *Neuron*, 94(2):375–387.e7.

Hart, P. E., Nilsson, N. J., and Raphael, B. (1968). A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107.

Hartley, T., Burgess, N., Lever, C., Cacucci, F., and Keefe, J. O. (2000). Modeling Place Fields in Terms of the Cortical Inputs to the Hippocampus. *Hippocampus*, 379:369–379.

Hartley, T., Trinkler, I., and Burgess, N. (2004). Geometric determinants of human spatial memory. *Cognition*, 94(1):39–75.

Heffner, R. S. and Heffner, H. E. (1992). Visual factors in sound localization in mammals. *The Journal of Comparative Neurology*, 317(3):219–232.

Hinman, J. R., Chapman, G. W., and Hasselmo, M. E. (2019). Neuronal representation of environmental boundaries in egocentric coordinates. *Nature Communications*, 10(1):1–8.

Hsu, F. H. (1999). IBM's Deep Blue chess grandmaster chips. *IEEE Micro*, 19(2):70–81.

Huxter, J., Burgess, N., and O'Keefe, J. (2003). Independent rate and temporal coding in hippocampal pyramidal cells. *Nature*, 425(6960):828–832.

Ismakov, R., Barak, O., Jeffery, K., and Derdikman, D. (2017). Grid Cells Encode Local Positional Information. *Current Biology*, 27(15):2337–2343.e3.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. Technical Report 430.

Killcross, S. and Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral cortex*, 13(4):400–408.

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(6029).

Kolling, N., Wittmann, M. K., Behrens, T. E., Boorman, E. D., Mars, R. B., and Rushworth, M. F. (2016). Value, search, persistence and model updating in anterior cingulate cortex.

Kringelbach, M. L. (2005). The human orbitofrontal cortex: Linking reward to hedonic experience.

Kropff, E., Carmichael, J. E., Moser, M. B., and Moser, E. I. (2015). Speed cells in the medial entorhinal cortex. *Nature*, 523(7561):419–424.

Krupic, J., Bauza, M., Burton, S., Barry, C., and O'Keefe, J. (2015). Grid cell symmetry is shaped by environmental geometry. *Nature*, 518(7538):232–235.

Lever, C., Burgess, N., Cacucci, F., Hartley, T., and O'Keefe, J. (2002). What can the hippocampal representation of environmental geometry tell us about Hebbian learning? *Biological Cybernetics*, 87(5-6):356–372.

Lever, C., Burton, S., Jeewajee, A., O'Keefe, J., and Burgess, N. (2009). Boundary Vector Cells in the Subiculum of the Hippocampal Formation. *Journal of Neuroscience*, 29(31):9771–9777.

Maidenbaum, S., Miller, J., Stein, J. M., and Jacobs, J. (2018). Grid-like hexadirectional modulation of human entorhinal theta oscillations. *Proceedings of the National Academy of Sciences*, 115(42):10798–10803.

Mcnaughton, B. L., Barnes, C. A., and O'keefe, J. (1983). The Contributions of Position, Direction, and Velocity to Single Unit Activity in the Hippocampus of Freely-moving Rats. Technical report.

Mehta, M., M.C. Quirk, and Wilson, M. (2000). Experience-Dependent Asymmetric Shape of Hippocampal Receptive Fields. *Neuron*, 25:707–715.

Miller, K. J., Botvinick, M. M., and Brody, C. D. (2017). Dorsal hippocampus contributes to model-based planning. *Nature Neuroscience*, 20(9):1269–1276.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., and Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9):680–692.

Morris, R. and Ward, G. (2004). *The cognitive psychology of planning*. Psychology Press.

Morris, R. G. M., Garrud, P., Rawlins, J. N. P., and O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature*, 297(5868):681–683.

Moser, E. I. (2011). The multi-laned hippocampus.

Muller, R. and Kubie, J. (1987). The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *The Journal of Neuroscience*, 7(7):1951–1968.

Muller, R., Kubie, J., and Ranck, J. (1987). Spatial firing patterns of hippocampal complex-spike cells in a fixed environment. *The Journal of Neuroscience*, 7(7):1935–1950.

Muller, R. U., Bostock, E., Taube, J. S., and Kubie, J. L. (1994). On the directional firing properties of hippocampal place cells. *Journal of Neuroscience*, 14(12):7235–7251.

Oja, E. (1982). A Simplified Neuron Model as a Principal Component Analyzer. Technical report.

O'Keefe, J. (1976). Place units in the hippocampus of the freely moving rat. *Experimental Neurology*, 51(1):78–109.

O'Keefe, J. and Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons.

O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 34(1):171–175.

O'Keefe, J. and Nadel, L. (1978). *The hippocampus as a cognitive map.* Oxford: Clarendon Press.

Packard, M. G. and McGaugh, J. L. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory*, 65(1):65–72.

Pearce, J. M., Roberts, A. D., and Good, M. (1998). Hippocampal lesions disrupt navigation based on cognitive maps but not heading vectors. *Nature*, 396(6706):75–77.

Poulter, S., Hartley, T., and Lever, C. (2018). The Neurobiology of Mammalian Navigation. *Current Biology*, 28(17):R1023–R1042.

Quirk, G. J., Muller, R. U., and Kubie, J. L. (1990). The firing of hippocampal place cells in the dark depends on the rat's recent experience. *Journal of Neuroscience*, 10(6):2008–2017.

Ramon y Cajal, S. (1911). Histologie du système nerveux de l'homme et des vertébrés. *Maloine, Paris*, 2:153–173.

Raudies, F. and Hasselmo, M. E. (2012). Modeling Boundary Vector Cell Firing Given Optic Flow as a Cue. *PLoS Comput Biol*, 8(6):1002553.

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., and Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational Biology*, 13(9).

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(5774):758–762.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, 275(5306):1593–1599.

Shenhav, A., Botvinick, M. M., and Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function.

Simon, D. A. and Daw, N. D. (2011). Neural Correlates of Forward Planning in a Spatial Decision Task in Humans. *Journal of Neuroscience*, 31(14):5526–5539.

Solstad, T., Boccara, C. N., Kropff, E., Moser, M. B., and Moser, E. I. (2008). Representation of geometric borders in the entorhinal cortex. *Science*, 322(5909):1865–1868.

Sorscher, B., Mel, G. C., Ganguli, S., and Ocko, S. A. (2019). A unified theory for the origin of grid cells through the lens of pattern formation. Technical Report NeurIPS.

Stachenfeld, K. L., Botvinick, M. M., and Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*.

Stensola, H., Stensola, T., Solstad, T., FrØland, K., Moser, M. B., and Moser, E. I. (2012). The entorhinal grid map is discretized.

Stensola, T., Stensola, H., Moser, M. B., and Moser, E. I. (2015). Shearing-induced asymmetry in entorhinal grid cells. *Nature*, 518(7538):207–212.

Strange, B. A., Witter, M. P., Lein, E. S., and Moser, E. I. (2014). Functional organization of the hippocampal longitudinal axis.

Sul, J. H., Kim, H., Huh, N., Lee, D., and Jung, M. W. (2010). Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. *Neuron*, 66(3):449–460.

Sun, Y., Jin, S., Lin, X., Chen, L., Qiao, X., Jiang, L., Zhou, P., Johnston, K. G., Golshani, P., Nie, Q., Holmes, T. C., Nitz, D. A., and Xu, X. (2019). CA1-projecting subiculum neurons facilitate object–place learning. *Nature Neuroscience*.

Sun, Y., Nguyen, A. Q., Nguyen, J. P., Le, L., Saur, D., Choi, J., Callaway, E. M., and Xu, X. (2014). Cell-type-specific circuit connectivity of hippocampal CA1 revealed through cre-dependent rabies tracing. *Cell Reports*, 7(1):269–280.

Sun, Y., Nitz, D. A., Holmes, T. C., and Xu, X. (2018). Opposing and complementary topographic connectivity gradients revealed by quantitative analysis of canonical and noncanonical hippocampal CA1 inputs. *eNeuro*, 5(1).

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition.

Taube, J. S., Muller, R. U., and Ranck, J. B. (1990). Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 10(2):420–35.

Tolman, E. C. (1948). Cognitive maps in rats and men.

Tolman, E. C. and Honzik, C. H. (1930). Introduction and removal of reward, and maze performance in rats. *University of California Publications in Psychology*, 4:257–275.

Tremblay, L. and Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature*, 398(6729):704–708.

Ulanovsky, N. and Moss, C. F. (2007). Hippocampal cellular and network activity in freely moving echolocating bats. *Nature Neuroscience*, 10(2):224–233.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-Learning. *Machine Learning*, 8(3-4):279–292.

Yartsev, M. M. and Ulanovsky, N. (2013). Representation of three-dimensional space in the hippocampus of flying bats. *Science*, 340(6130):367–372.

Yartsev, M. M., Witter, M. P., and Ulanovsky, N. (2011). Grid cells without theta oscillations in the entorhinal cortex of bats. *Nature*, 479(7371):103–107.

Zhang, J., Springenberg, J. T., Boedecker, J., and Burgard, W. (2017). Deep reinforcement learning with successor features for navigation across similar environments. In *IEEE International Conference on Intelligent Robots and Systems*, volume 2017-Septe, pages 2371–2378. Institute of Electrical and Electronics Engineers Inc.

Zhang, S. J., Ye, J., Couey, J. J., Witter, M., Moser, E. I., and Moser, M. B. (2014). Functional connectivity of the entorhinal - Hippocampal space circuit.