Author:
Roscow, Emma L

Title:
Modelling reward-dependent replay of memory coordinated across hippocampus and nucleus accumbens

Modelling reward-dependent replay of memory coordinated across hippocampus and nucleus accumbens

Emma Louise Roscow

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Life Sciences

September 2019

Word count: 47,376

Abstract

During sleep and rest, the brain engages in internally generated sequences of activity which reflect encoding of recent experiences. This replay of neural activity is believed to promote memory consolidation, preferentially reinforcing the memories encoded by the replayed activity. It is thought that preferential replay of some activity can optimise the processing and retention of new information to optimise future behaviour, but it remains unclear which experiences are prioritised for replay, or which features of an awake experience influence replay prioritisation. Replay depends on the hippocampus, a brain structure heavily involved in forming new memories, and recruits a wide range of other brain areas to enable systems-level consolidation coordinated across the brain.

This thesis presents a combination of computational modelling and in vivo behavioural and electrophysiological experiments used to investigate how reward and non-reward experiences influence replay. I developed and implemented a maze-based reinforcement learning task with stochastic rewards, in which both rewarded and unrewarded trials are informative for learning.

Computational modelling based on a reinforcement learning framework showed that biasing replay by reward or reward-prediction error can enhance learning, but replay of a range of trials (rewarded and unrewarded) is necessary. Modelling of rats' behaviour on the same task suggested that they preferentially replayed experiences which generated high reward-prediction errors between training sessions. Preliminary multi-unit recordings made from the hippocampus, which is heavily implicated in replay, and the nucleus accumbens, which responds to reward and receives input from the hippocampus, suggests that neural activity encoding spatial and reward information in these two structures is replayed during post-task rest. This is a likely mechanism by which reinforcement learning can occur after behaviour has taken place.

The work presented in this thesis extends the understanding of how reward influences learning, memory consolidation and replay, particularly in offering evidence for the biasing effect of reward-prediction error.

Acknowledgements

First of all, my thanks and appreciation go to my supervisors, Matt Jones and Nathan Lepora, for the intellectual freedom and encouragement they have given me throughout my PhD in pursuing the project of my choice. I am grateful for their support, time and constant availability throughout the last four years.

I am also grateful to members of the Jones lab for advice and support, as well as the Neural Dynamics group and countless other people who have offered wisdom, insightful questions and helpful comments at all stages of this research. Particular thanks to Aleks, who has been characteristically more generous with his time than he was obligated to be. Countless non-academic staff provided services without which this work would not have been possible, not least the university's Advanced Computing Research Centre which granted access to high-performance computing facilities, the Animal Services Unit which took excellent care of husbandry for the rats, and the Wellcome Trust which granted me the opportunity and funding to embark on research in a fascinating area.

And, of course, I am indebted to the rats.

Finally, special thanks go to Mickaël for his patience, forbearance, constant support and tales of Gui-Gui which kept me going.

Author’s declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED …… DATE ……

Contents

List of Figures

List of Tables

Chapter 1: Introduction

What does the brain do at rest? During activity, the brain's principal function is to process incoming sensory information, interpret, infer, predict, and produce appropriate actions. When the body is not active, although the brain is ostensibly unconstrained by these duties, it remains active nonetheless, producing spontaneous spiking activity dominated by its own experience- and state-dependent dynamics. What happens in the brain during sleep reflects prior waking experience and, in turn, activity during sleep contributes to the cognitive processing that instructs future behaviour.

The aim of this thesis is to understand how neural processing at one time influences, predicts and shapes neural processing later on: across the time course of learning, and over the sleep-wake cycle. The focus is on the interrelationships between neural encoding of memories that bind location and reward, which are important for successful spatial reinforcement learning, and to test the hypothesis that consolidation of those memories are supported by interactions between the hippocampus (known to be involved in processing spatial information) and the nucleus accumbens (involved in processing reward).

This thesis is divided into seven chapters. The current chapter gives a conceptual overview of the research contained within the thesis and the literature that informed the motivations for this project. The background and methods for each of the five results chapters are reasonably distinct, so a more thorough review of the literature pertaining to each chapter – as well as the details of the methods used – are given in the introduction to each chapter. In this sense, the literature review is spread throughout the thesis; the current chapter gives a unified overview of the motivation for the research which is expanded on in each subsequent chapter.

1.1. Place cells

One archetypal area of research in which internally generated spiking activity has proved to be of interest is in the activity of hippocampal place cells. Place cells in the hippocampus were first observed nearly 40 years ago in behaving rats, with the distinctive characteristic of exhibiting spatial receptive fields: they increased their firing rate significantly when the rat was in a particular location (O'Keefe & Dostrovsky 1971), offering a read-out of the rat's position based on the firing rates of the cells. Advances in technology which permit simultaneous recording of tens (Jun et al., 2017; Wilson & McNaughton, 1993) to hundreds (Villette et al., 2015) of place cells have revealed complex spatial coding properties.

Place cells have proved to be a useful instrument for probing the neural correlates of behaviour, because of both the relative ease of mapping their activity to an animal's ongoing behaviour and the capacity to manipulate that behaviour through the use of mazes, linear tracks and other kinds of spatial structure imposed on animals' movements with good ethological validity. Anatomically, the hippocampus is an easily accessible brain structure to record from in vivo, and its distinct cellular structure makes it easily identifiable in vitro, which has encouraged a large body of experimental work on the hippocampus to flourish over decades; the detailed known hippocampal circuit anatomy is another advantage for studying place cells. Place fields form a distinct map for every environment, so as a result, they have been the focus of a large body of research on learning and memory in which animals are introduced to new environments, new spatial tasks or other spatial manipulations. By necessity, place fields tread the fragile line between plasticity (the requirement to flexibly encode new environments as they are encountered, or update their map to reflect environmental changes) and stability (the requirement to persist for as long as the animal navigates the environment – potentially its whole life). Place fields form quickly, within five minutes of exposure to an environment (Frank et al., 2004), but remap when an environment grows (Rich et al., 2014), when it merges with another previously distinct environment (Bostock et al., 1991), or to cluster around areas of importance like reward locations (Hollup et al., 2001). Despite these dynamic qualities, some element of place coding remains stable for as long as experimental constraints have allowed – weeks – and likely longer (Ziv et al., 2013).

Externally and internally generated sequences are crucial to how place cells function. By the nature of locomotion, place cells are activated in sequence over a period of seconds as an animal moves through its environment; but these sequences are also apparent on a finer timescale. During behaviour the hippocampus is dominated by theta oscillations, cycles of widespread, synchronised excitation alternating with inhibition that manifest at a frequency of roughly 6-8 Hz (Buzsáki, 2002). The spiking activity of place cells becomes locked to these oscillations, spiking during a regular window in each cycle relative to the phase of the cycle, but they precess over time, firing slightly earlier in each theta cycle as the animal moves through the place field (fig. 1.1). Place fields overlap, which means that more than one place cell is active

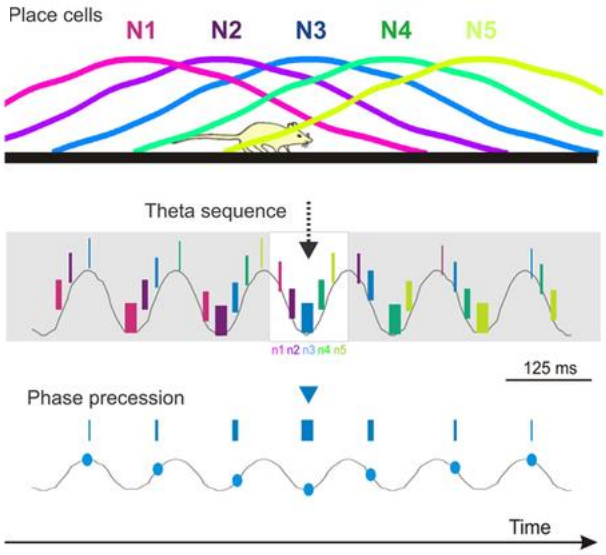

Figure 1.1. Theta-frequency phase precession. Top: the firing rates of five place cells (N1-N5) in the hippocampus as the animal moves through space, activating each cell in turn. These cells fire in sequence on a behaviourally-relevant timescale. Middle: spiking activity of the same cells relative to theta oscillations in the hippocampal LFP; thicker bars indicate stronger firing. Entrainment to theta oscillations results in similar sequences of place cell activity within each theta cycle which reflects behaviour on a longer timescale. Bottom: a cell will fire late in the theta cycle as the animal enters the place field, at the trough of a cycle when the animal is in the middle of the place field, and early in the theta cycle as the animal exits the place field, giving rise to the term "phase precession". (Adapted from Dragoi (2013).)

in a given theta cycle; the result is that the place cell sequence which plays out over seconds as the animal moves is also played out, on a faster timescale, within each theta cycle (Foster & Wilson, 2007; O'Keefe & Recce, 1993; Skaggs et al., 1996).

These theta cycles have the property of encoding a portion of the path already taken, as well as predicting a portion of the path yet to be taken, offering a kind of instantaneous link through time (Gupta et al., 2012) which may promote plasticity between the active cells. This is particularly because these sped-up theta sequences promote shorter latencies between the spikes of consecutively active place cells than the behavioural timescale would allow, ensuring spike latencies which are conducive to spike-time-dependent plasticity (Sato & Yamaguchi, 2003). What is striking is that the same place cell sequences which play out in theta cycles are also recapitulated when the animal is not moving at all: before the locomotion begins (i.e. anticipating future behaviour; Ólafsdóttir et al., 2015; Pfeiffer & Foster, 2013) and after the locomotion ends (i.e. recalling past behaviour; Foster & Wilson, 2006). At these times, place cells fire in sequences that resemble the activity during running on the to-be-taken or just-taken path, a phenomenon known as replay. Unlike theta sequences, which are evoked as a result of sensorimotor engagement with the environment,

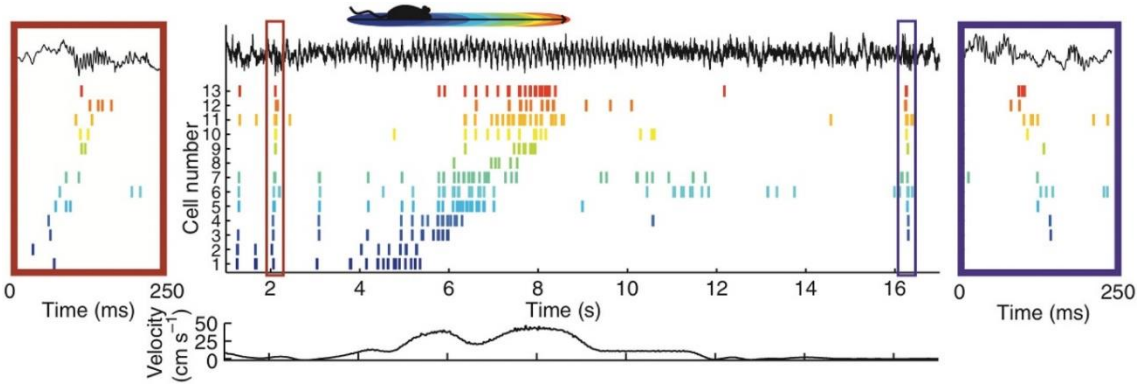

Figure 1.2. Replay of hippocampal place cell sequences. As a rat travels through its environment, it passes through the place fields of a set of hippocampal pyramidal cells, activating them in sequence (spikes shown as coloured ticks, one row per cell). This is accompanied by hippocampal LFP dominated by theta power (black trace, above). When the animal is at rest, brief, transient activation of the same cells is apparent, in a similar sequence but compressed in time (red and blue boxes). This replay can occur in the forwards direction (red box), or in the opposite direction (blue box). (Taken from Kay and Frank (2019), adapted from Diba and Buzsáki (2007).)

replay sequences are spontaneously generated by the hippocampal network during times when the animal is behaviourally disengaged from its environment (Kay & Frank, 2019).

The close association between place cell activity during replay ("offline") and activity during behaviour ("online") invites theories about how replay might influence behaviour by supporting or contributing to cognition. Its involvement has been hypothesised in a number of cognitive processes: foremost is memory consolidation, consistent with the observation that (like online theta sequences) replayed sequences are compressed in time to timescales which promote long-term potentiation (LTP), strengthening the synapses between cells engaged in the replay event (King et al. 1999). Replaying place cell sequences, especially ones relating to a new environment or changes to a familiar environment, might stabilise place maps by potentiating the synapses of adjacent place cells (Dupret et al., 2010).

Replay of upcoming sequences of place cell activity (sometimes called preplay) has been suggested to enable planning or decision-making related to ongoing behaviour: at choice points on a maze, place cell sequences appear to encode multiple possible forwards paths in alternation (Johnson & Redish, 2007; van der Meer et al., 2010), and the content of replay relates to changing priorities in goal-directed navigation (Carey et al., 2019), perhaps enabling an evaluation of possible choices before selecting one. A related idea is that replay might be associated with spatial working memory, allowing locations visited in the recent past to be held in mind while choosing the next path to take, in complex navigational tasks which require locations to be visited in a particular order (Jadhav et al., 2012). The constant playing out of sequences reflecting current, past, and future behaviour is a robust and widely reported finding, which suggests a tight link to cognitive processes.

Such replay or preplay sequences are observed not only during active behaviour, evaluation and decision-making, but also during periods of extended rest. Bursts of place cell activity encoding paths taken by an animal are observed for periods extending sometimes hours after the behaviour has occurred (Giri et al., 2019), beyond what can be expected by chance, and encoding locations or environments remote from the animal's current position. This replay is seen particularly strongly during slow-wave sleep, and more debatably in rapid-eye-movement (REM) sleep, and is thought to underlie the observation of sleep-dependent memory consolidation which has been noted for a century or more (Jenkins & Dallenbach, 1924).

Replay during extended rest and sleep appears to also have an important function in cognitive processes. In particular, it promotes goal-directed behaviour during subsequent training sessions on the same task, suggesting a role in reinforcement learning (Cazé et al., 2018; Johnson & Redish, 2005). This is evident both from experiments which interrupt replay processes, impairing the subsequent ability to identify locations at which reward is available (Ego-Stengel & Wilson, 2010; Girardeau et al., 2009), and experiments which stimulate the activity of place cells during sleep, promoting a preference for visiting the reinforced place fields (de Lavilléon et al., 2015). The additional, replayed activity appears to further reinforce rewarded actions, and this may rely on the recruitment of other brain areas into the replay event which play a role in reward processing, including the prefrontal cortex (Euston et al., 2007; Peyrache et al., 2009) and striatum (Pennartz et al., 2004). Indeed, replay of task-relevant activity in brain structures beyond the hippocampus is observed coincident with replay of place cell activity, which means that replay processes go beyond navigation and spatial maps to more wide-ranging cognitive applications.

To some extent, all of these processes can be viewed as forms of memory consolidation, processing or reprocessing past experience. The hippocampus has long been implicated in the formation of new memories, and replay offers a plausible mechanism by which new memories might be stabilised and retained for guiding future behaviour. But the hippocampus's involvement in retrieval of distant memories is much weaker, especially for non-spatial memories. This time-dependent manner of hippocampal involvement in memory retrieval suggests that the process of consolidation happens at a systems level, reorganising the neural trace of a memory from its initial configuration in the hippocampus to a more extra-hippocampal representation. The role of hippocampal replay in systems-level memory consolidation is considered next.

1.2. Replay and memory consolidation

Observations of patients with amnesia resulting from lesions to the hippocampus have long suggested that the hippocampus is vital to memory function. Hippocampal damage selectively impairs anterograde rather than retrograde amnesia, a pattern which points to the importance of the hippocampus for forming new memories but not necessary for recalling existing, more temporally remote memories (Scoville & Milner 1957). This phenomenon has been corroborated more recently with in vivo recordings (Frankland & Bontempi, 2005), along with in vitro studies which show changes in the synapses of recently active cells in the hippocampus over a period of hours after initial learning, which points to a time-dependent process of memory consolidation (Dudai, 2004; Frankland & Bontempi, 2005). Engram cells which are active during a novel experience appear quickly in both the hippocampus and the prefrontal cortex; early in learning, disrupting the activity of hippocampal engram cells impairs retrieval of the recent memory, but after some days the hippocampal input becomes unnecessary and prefrontal cortical engram cells take over the mnemonic role (Kitamura et al., 2017).

This time-dependent process of initial encoding followed by later consolidation has been suggested as the solution to a major dilemma: how does the brain mediate the difficult balance between plasticity and stability? The complementary learning systems theory (McClelland et al.,1995) proposes that the brain takes advantage of two modes of learning: a fast system, implemented by the hippocampus, which forms representations quickly and sparsely but has limited capacity for storage; and a slow system, implemented by the cortex, which has a lower learning rate to take advantage of the statistical regularities between similar experiences and encodes experiences in a more distributed way throughout large populations (Kumaran et al., 2016). The fast, sparse encoding enables pattern separation which is beneficial for one-short learning or retention of individual episodes of experience, but is not suited to generalising across episodes. In contrast, the slow learner can integrate multiple episodes by slightly adjusting its connection weights to encode similarities between them, but takes many examples to achieve this. Sleep may be important for this process because it provides an opportunity for additional training, when internally generated patterns of activity can dominate and induce synaptic changes without interference from incoming sensory information. The brain exhibits differences in global state in many respects: concentrations of various neurotransmitters are different (Samanta et al., 2020), local field potential is dominated by different oscillations, and many neurons show different patterns of firing (Mahon et al., 2006; Miller et al., 1983; Vyazovskiy et al., 2009), indicating that unique neural processes might be undertaken during sleep that are distinct from those during wake.

Many aspects of brain function are modulated by the sleep-wake cycle which have implications for learning and memory. Synaptic plasticity in the cortex increases throughout the hours of wake, evident from increased expression of plasticity-related genes, larger synapses (de Vivo et al., 2017), and increased firing

rates, which subside following a period of sleep. Boundless plasticity would eventually lead to saturation in firing rates, so this sleep-dependent homeostasis is vital for normal functioning. High-firing neurons reduce their firing rate and low-firing neurons increase it, leading to a reorganisation of cortical circuits which might offer the ideal physiological conditions for hippocampus-lead memory consolidation to take place (Born & Wilhelm, 2012). Indeed, sleep has been found to improve not only memory consolidation, but other functions as well which suggest a process of generalising from individual episodes or extracting statistical regularities (Stickgold & Walker, 2013). Insight, a sudden acquisition of latent knowledge, has been reported to be promoted by a period of sleep (Lewis et al., 2018; Wagner et al. 2004), as has false recall of words which are semantically similar to words which have been experienced (Diekelmann et al., 2010). Systems-level replay of activity across multiple brain areas might underlie these processes.

Merely teaching a slow learner using a fast learner does not solve the stability-plasticity problem, however. New information must be integrated with, not replace, existing structures, in order to ensure that the neural representations remain robust. This was illustrated in the example of the "penguin problem" (McClelland et al., 1995), in which a connectionist network which had learned about characteristics of birds and fish was faced with the challenge of incorporating the atypical, flightless, aquatic penguin into its representations. The network was capable of rapidly learning to categorising the penguin, but connection weights were adjusted so dramatically that this ability came at the expense of its correct representations of other birds. To preserve the existing representations, the penguin training set had to be interleaved with examples of older, already-acquired items. This suggests another function for hippocampal replay: complementary learning systems involving a fast learner and a slow learner can be used to interleave new and old information to avoid catastrophic forgetting.

Connectionist and other artificial neural network models such as this have proved useful for demonstrating many of the constraints and challenges of learning (Thomas & McClelland, 2008). As the fields of machine learning and artificial intelligence have grown, artificial agents which learn complicated rules based on few principles have often come up against similar problems to those outlined by theories such as the complementary learning systems theory. The mission of building intelligent systems through self-organised learning can be seen as reverse-engineering the brain, which offers the opportunity to investigate what functions particular neural phenomena might have for cognitive processes (Hassabis et al. 2017; Marblestone et al., 2016; Wang et al. 2018). In particular, further evidence for the value of interleaving training data has come from artificial neural networks with complicated architectures but simple learning rules, known as deep neural networks, and the machine learning field has focused efforts to optimise this interleaving.

1.3. Memory consolidation in machine learning

Deep neural networks, artificial neural networks composed of several layers of feedforward units, have shown remarkable learning abilities (Mnih et al. 2013; Silver et al. 2016), and face many of the same challenges as biological intelligent agents. Loosely modelled on the structure of the cortex, each layer is represented by a vector of values (abstractly equivalent to the firing rates of a population of neurons), which are propagated to the next layer, multiplied by connection weights. These weights are simplified representations of synaptic connections which are randomised prior to learning and slowly adjusted over the course of learning: the final output of the network generates some error which is backpropagated to update the weights closer towards producing a smaller error, effectively implementing stochastic gradient descent. Given enough training, a deep neural network approximates an optimisation function.

With no anatomical "hard-wiring" or intrinsic biases towards any particular type of computation, deep neural networks rely entirely on training data to adjust connection weights between units in successive layers to achieve goals. Often, as the agent learns to adapt its behaviour towards successful action choices, it biases its own training data towards the narrow range of what it "chooses" to experience, overfitting its parameters to a limited set of experiences. The effect of overfitting is to cause catastrophic forgetting of earlier learning which is no longer being enacted, hampering performance overall. As with the penguin problem, learning has been found to vastly improve when this temporal correlation is broken by interleaving ongoing, online training with samples taken from a memory buffer, a technique known as experience replay (Lin 1992; Schaul et al. 2015). This can afford the agent much more flexibility, not only learning current goals more quickly (Andrychowicz et al. 2017) but using knowledge gained from past goals to inform future goals (Rolnick et al. 2018; Shin et al. 2017) for better long-term transfer of knowledge from one domain to another.

When the technique was first developed, samples were selected uniformly from the memory buffer, but attempts in recent years to prioritise some experiences over others can further optimise deep reinforcement learning. If some experiences can be identified as more useful or instructive for training, they can be selectively or preferentially replayed to the agent to bias its learning in potentially useful ways (Moore & Atkeson, 1993; Schaul et al., 2015; Kumaran et al., 2016). In particular, inspired by findings in neuroscience about the effect of reward on hippocampal replay, samples in the memory buffer can be prioritised for replay according to the degree to which they elicit a high temporal-difference error (a measure of surprise about the outcome; Schaul et al., 2015). High temporal-difference errors indicate room for improvement, so these samples are the most informative for further offline training. The best method for prioritising is highly task-dependent (Liu & Zou, 2018), but at least for some tasks prioritised replay is one of the most successful techniques in deep reinforcement learning (Hessel et al., 2018)

An additional benefit of training from experience replay is that it makes more efficient use of real-world, online training data by reprocessing them. In the simplest case, this means simply re-sampling from the memory buffer, with policies for prioritisation that protect the system from catastrophic forgetting. Taking current goals into account when prioritising replay can further enhance learning by promoting samples which are the most relevant to current objectives. Taken to the extreme, whole models of the environment can be derived from limited online training, from which the system can generate new, hypothetical training data to sample from in ways more varied and less biased than sampling directly from the memory buffer when the buffer is small (Robins, 1995; Bruce et al. 2017; Shin et al., 2017). The benefit is that learning can converge with only a small amount of interaction with the environment, even to the point of affording one-shot learning from a single experience in the environment (Bruce et al. 2017).

Superficially, these replay algorithms resemble observations about hippocampal replay: replay of trajectories to or from an animal's current location are interleaved with replay of locations further away (Gupta et al., 2010), rewarded trajectories are replayed preferentially to unrewarded trajectories (Singer & Frank, 2009), and replay boosts learning as if making additional use of a limited training set (Ego-Stengel & Wilson, 2009). The innovation of generating models from which to sample and replay hypothesised experiences prompts the question of whether this, too, is a feature of hippocampal replay which enhances learning, a theory which is considered next.

1.4. Hippocampus as a prediction generator

There is great variation in the characteristics of internally generated place cell sequences. Replay events have been reported spanning sleep, extended periods of wakeful rest (Skaggs & McNaughton, 1996), and brief moments of rest (Foster & Wilson, 2006) in which past activity is replayed. This is in addition to momentary activity during both theta epochs (Johnson & Redish, 2007) and ripples which predicts upcoming trajectories, as well as longer periods of rest and sleep in which activity occurs that resembles place cell sequences not yet experienced (Dragoi & Tonegawa, 2011; Ólafsdóttir et al., 2015). Replay events can reflect locations local to or remote from the animal's current position (Gupta et al., 2010), and play in the forward direction (i.e. reflecting the direction of the path the animal has taken) or backwards in reverse (Foster & Wilson, 2006). It remains unclear whether these represent functionally distinct "types" of replay or whether a common mechanism underlies all of these sequences (Joo & Frank, 2018; Mattar & Daw, 2018), but it is tempting to associate replay after experience with a process of memory consolidation, and replay (or preplay) before experience to planning. In any case, such sequences do appear to relate to

ongoing behaviour and task demands: greater reactivation during ripples is associated with better spatial learning performance on the subsequent trial (Singer et al., 2013), with some suggestion that upcoming trajectories are over-represented during these ripples; although in a task involving discounting of a previously valued choice, replay is biased towards the discounted and unchosen option (Carey et al., 2019). Replay content is more closely associated with current location during brief periods of immobility than extended rest (Ólafsdóttir et al., 2017), and the balance of forwards and backwards replay is modulated by reward (Ambrose et al., 2016) and sleep-wake state (Wikenheiser & Redish, 2013). This dynamic nature of replay in relation to behaviour suggests an integral role in cognition, although definitively what aspect or aspects of cognition depend on replay is an ongoing question.

The observation of predictive coding in place cell sequences prompts the idea that the hippocampus produces generative models, building a map of the environment through experience and sampling from it in a flexible, task-dependent way to enable cognition (Chersi & Pezzulo, 2012; Pezzulo et al., 2017). The idea of a generative model hippocampus is consistent with findings from patients with hippocampal damage who tend to have difficulty imagining future or hypothetical episodes (Hassabis et al., 2007; Ólafsdóttir et al., 2018; Tulving, 1985). During decision-making at a choice point on a maze, for example, the hippocampus appears to generate forward sweeps of place cell activity encoding one possible forward path followed by another (Johnson & Redish, 2007; Redish, 2016; Tolman 1948). This tends to be accompanied by head movements orienting towards the corresponding paths, known as vicarious trial-and-error (VTE), as well as activity in the reward-responsive nucleus accumbens brain structure (Stott & Redish, 2014) which suggests a mechanism of considering and evaluating actions and outcomes before committing to one.

Replay sequences do not necessarily consist of faithful accounts of real experiences, further suggesting that the hippocampus engages in active simulation rather than passive replay: as well as sequences that are in the reverse order from what the animal has experienced (Diba & Buzsáki, 2007; Foster & Wilson, 2006), sequences are observed which are decodable to trajectories that have never been taken by the animal (Gupta et al., 2010), or a reorganisation of the temporal order in which they were experienced (Liu et al., 2019). This may result from random activity amongst place cells in the CA3 hippocampal subregion, which perform pattern-completion owing to their recurrent connections to form an attractor network (Shen & McNaughton, 1996), so that random activation of one part of the cognitive map triggers activation of another – regardless of the experiential relationship between them. In CA3, replay in either direction may promote synaptic plasticity between place cells, as symmetric spike-time-dependent plasticity (STDP) has been found which is indifferent to the temporal order of pre- and post-synaptic spikes (Mishra et al., 2016).

1.5. Synaptic plasticity and dopamine

In the brain, the association between reward, motivation, prediction errors and plasticity relies critically on dopamine, which acts at synapses to influence plasticity (Gerstner et al., 2018). The hippocampus receives two sources of dopamine, from locus coeruleus and ventral tegmental area (VTA; McNamara et al., 2014; McNamara and Dupret, 2017) with observable effects on learning and plasticity. Dopamine increases the excitability of cells in hippocampus (Otmakhova & Lisman, 1996) and induces synaptic plasticity, the cellular basis for memory formation, hours after activation (Frey et al., 1990), coinciding with the time course of replay. Plasticity among place cells is modulated by reward-related processes, as CA1 place fields remap in response to reward, clustering around goal locations (Dupret et al., 2010). This is likely to be dopamine-dependent: administration of a dopamine receptor agonist during spatial exploration enhances place cell stability while antagonists reduce stability (Kentros et al., 2004), and stimulation of dopaminergic VTA terminals in CA1 causes increased coactivation during subsequent rest (McNamara et al., 2014). The effect of dopaminergic activity on behaviour is apparent from a study in which the dopaminergic medial forebrain bundle was stimulated whenever a particular place cell was activated during sleep; subsequently animals showed a preference for the location which overlapped with the reinforced place field (de Lavilléon et al., 2015).

In theory, then, if a mesolimbic dopaminergic circuit is preferentially reactivated during sleep it could selectively promote plasticity to reinforce particular behaviours, effectively boosting reinforcement learning offline. There is evidence that similar mechanisms are at work during sleep: a similar network to the one which appears to generate place cell sequences at choice points and evaluate the likely outcome also engages in replay during sleep (Lansink et al. 2008, 2009). Replay has been observed in the nucleus accumbens, which has been associated with generating reward prediction signals, and the VTA, which signals reward-prediction error (Gomperts et al., 2015; Valdés et al., 2015). The hippocampus, nucleus accumbens and VTA form a functional loop, whereby spatial and novelty information is transmitted from the hippocampus to the accumbens, which disinhibits the VTA (via the ventral pallidum; Lisman & Grace, 2005). This disinhibition promotes bursty, phasic firing which triggers dopaminergic release in a large number of brain structures to which VTA projects, including back to the hippocampus and accumbens, which in turn enhances long-term potentiation. The pathway from CA1 to VTA directly affects spatial reinforcement learning (Esmaeili et al., 2012), so dopamine-mediated plasticity in this circuit during sleep and rest could further promote reinforcement learning.

Consistent with this proposal, dopamine has been shown to affect neuronal excitability and plasticity not only within the hippocampus but also at hippocampal-accumbens synapses. Dopamine receptors in the accumbens modulate hippocampus-evoked responses, dynamically controlling the degree to which hippocampal inputs are transmitted to the accumbens (Goto & Grace, 2005). Dopamine enhances long-

term potentiation at hippocampal-accumbens synapses too (Schotanus & Chergui, 2008), so the effect of reward-related dopaminergic release is to promote the reorganisation of assemblies in hippocampus and at the connections between hippocampus and accumbens. Although the effect of dopamine on these synapses has not been investigated directly in relation to replay, accumbens cells which respond to reward are preferentially recruited into hippocampal replay processes (Lansink et al., 2009).

The behavioural and environmental correlates of dopamine release at these synapses might offer some clue as to the content of what gets replayed in this circuit. The signals that VTA sends to the accumbens are complex: individual VTA dopaminergic cells project their axon terminals over large distances in the accumbens, but these axon terminals are subject to the activation of a variety of presynaptic receptors that control local dopamine release. While the firing rate of many VTA cells encodes prediction errors (i.e. surprise, especially surprising reward), the resulting dopamine release is reported to correlate better with value or expected reward (Papageorgiou et al., 2016; Berke, 2018). Although an influence of reward on replay has been established, whether reward, reward-prediction errors, or expected value are better determinants of the activity that gets replayed is unknown, particularly because these elements are usually correlated in most reinforcement learning tasks.

Finally, the time course of when dopaminergic action takes effect on the replayed activity is a further consideration. Hippocampal replay has primarily been associated with the slow-wave period of sleep, when dopamine concentrations in the brain are at their lowest (Samanta et al., 2020), and VTA cells have been reported to take part in replay events during wake but not sleep (Gomperts et al., 2015), which suggests that the influence of dopamine on replay takes effect during behaviour, not during sleep replay itself. In contrast, replay within accumbens network has been reported during slow-wave sleep (Pennartz et al., 2004). If the actions of dopamine on these synapses is different during behaviour from during replay, the relationship between dopamine-mediated reward signals in the hippocampus-accumbens network, information processing during learning, plasticity, and replay is further obscured.

The involvement of dopamine and the recruitment of the hippocampus-accumbens-VTA loop offer a mechanism for reinforcement during behaviour, and the recruitment of this network during post-task sleep and rest suggest that it plays a role in biasing the consolidation of memories which involve this network (Gomperts et al., 2015). Although there is some evidence that VTA and accumbens preferentially engage in replay of salient or reward-related information, how this relates to learning, or whether it can bias learning is beneficial ways, is unclear.

1.6. Overview of this thesis

This PhD project begins with the question, how does replay benefit reinforcement learning? Previous work has identified the importance of hippocampal replay to spatial learning tasks, and preferential replay of some activity appears to bias subsequent behaviour. This is evident both from observations that activity associated with reward or aversive experiences is replayed preferentially to neutral experiences (Ambrose et al., 2016; Girardeau et al., 2017; Gomperts et al., 2015; Pfeiffer & Foster, 2013; Singer & Frank, 2009; Valdés et al., 2015; Wu et al., 2017), and manipulations in which the content of replay is biased (Bendor & Wilson, 2012; Schouten et al., 2017). However, a rigorous investigation of how experiences might be prioritised for replay to optimise learning is lacking. In particular, for reinforcement learning, experiences which drive prediction errors are the most instructive to learn from because they indicate a deficiency in the model or representation which produced the prediction. The difference between experiences with high salience compared to high prediction error is most apparent when outcomes are stochastic, as neutral outcomes (e.g. no reward) can be as informative for guiding behaviour as salient outcomes (e.g. reward). For this reason, I have developed a spatial reinforcement learning task with stochastic rewards to probe how biased replay might aid learning. Given the suggestions that replay is important for recruiting non-hippocampal brain areas into systems-level consolidation processes, that reward-prediction errors might be important for biasing replay, and the involvement of the nucleus accumbens in processing reward, reward prediction and reward-prediction error, simultaneous in vivo recordings were made from the hippocampus and nucleus accumbens. Recordings were made during the task as well as during rest before and afterwards, to measure experience-dependent changes in activity coordinated between the two areas. To fully characterise how reinforcement learning can be enhanced by replay on the task, and to generate testable predictions about biological replay, learning on the task was simulated and modelled using a reinforcement learning algorithm.

Chapter 2 introduces the spatial learning task, and presents results from a pilot study in which rats were trained on it, to validate and characterise the behaviour. Chapter 3 shows neural correlates of behaviour on the task, recorded from the hippocampus and nucleus accumbens in vivo, to confirm that these areas are engaged in the task. Chapter 4 investigates computationally how replay might enhance learning, by running simulations of the task with a number of variations of a reinforcement learning model with replay. Chapter 5 fits the parameters of the model to behaviour on the task to infer how replay influences learning. Chapter 6 relates the computational findings to neural activity by identifying offline hippocampal-accumbens activity and relating it to activity during the task. Finally, Chapter 7 brings these results together to consider what can be learned about how offline activity influences reinforcement learning. Methods and further details of the relevant literature are covered in each chapter respectively.

Chapter 2: Spatial reinforcement learning task with stochastic rewards

The bulk of this chapter has been published online as a pre-print (Roscow et al., 2019).

2.1. Introduction

Much adaptive behaviour requires integrating and transferring information through time. On short time scales, working memory requires maintenance of information over seconds to execute an appropriate action; on long times scales, integrating knowledge over days is necessary to inform behavioural policies. There is evidence that, in spatial memory tasks, upcoming actions and outcomes are planned or anticipated (Carey et al., 2019; Dragoi & Tonegawa, 2011; Johnson & Redish, 2007; Muenzinger, 1931; Ólafsdóttir et al., 2015; Pfeiffer & Foster, 2013; Redish, 2016), and past outcomes are consolidated and generalised (Carr et al., 2011; Diba & Buzsáki, 2007; Foster & Wilson, 2006; Gupta et al., 2010; Jackson et al., 2006). These activities require processing of information beyond immediate sensory experience, using internally-generated processes to link past, present and future events (Buhry et al., 2011; Mattar & Daw, 2018; Pezzulo et al., 2017; Wikenheiser & Redish, 2015). To probe how experiences are linked together through time to guide behaviour, I developed a novel behavioural task which requires integration of learned information about space, actions and rewards to make optimal decisions, which allowed analysis of behaviour (this chapter), computational mechanisms (Chapters 4 and 5) and electrophysiological dynamics (Chapters 3 and 6) to uncover how the brain learns from past experience.

Much of the history of the study of animal behaviour has focused on the influence of reward on future behaviour: when a good outcome reliably follows a particular action with a short delay, the action is repeated

more often (Staddon & Cerutti, 2003; Thorndike, 1898). This adaptation of behaviour is variously known as instrumental learning, operant conditioning or reinforcement learning, and requires assigning credit for the outcome to previous actions which may have taken place many seconds before (Minsky, 1961; Walsh & Anderson, 2011). With deterministic outcomes, such as arrival at a particular location which guarantees reward, minimal examples of the action-outcome pair are required for the association to be learned. With stochastic outcomes, when reward is not guaranteed but delivered probabilistically, the challenge is harder: feedback is sparser, as the action must be repeated more times to obtain the same outcome and obtain a reliable sampling distribution of actions and outcomes. When there are multiple alternative actions with different stochastic outcomes, the optimal behaviour requires keeping track of average outcomes and comparing them, integrating individual experiences through time to store an average (Gittins & Jones, 1979).

In rats learning spatial tasks, behaviour (both the motor aspect and the decision-making involved) appears to become more automated with increasing familiarity, a process which has neural correlates and cognitive implications as well as behavioural manifestations. This has been described as occurring in three distinct stages, starting with a deliberative decision-making process and ending with automated behaviour (Redish, 2016). The initial, goal-directed decision-making process is dependent on the hippocampus and also recruits other brain areas including cortex and thalamus (Ito et al., 2015), using past experience to evaluate the best course of action. The final, automated stage is less dependent on these structures and more dependent on the basal ganglia, which drives habit-based behaviours by releasing highly efficient action chains (Aldridge et al., 2004; Graybiel, 1998). The advantage of the automated stage is that it bypasses much of the cognitive processing employed by the hippocampus, so actions are faster; but this comes at the cost of flexible action selection in response to a changing or uncertain environment. Coinciding with strong hippocampal involvement during the deliberative state in rodents is vicarious trial-and-error (VTE), apparently a behavioural manifestation of deliberation: animals orient towards one possible path, before reorienting towards another prior to initiating a run, as if weighing up the options (Redish, 2016; Tolman, 1948). This has been cited as evidence of a model-based, explicit hippocampal cognitive map of the environment, to which the animal can refer to make predictions about possible behaviour (Tolman, 1948). During VTE, hippocampal sweeps – sequences of place cell activity – are seen in the hippocampus which encode the possible future trajectories (Johnson and Redish, 2007), recruiting orbitofrontal cortex and nucleus accumbens as part of the deliberative process (van der Meer et al., 2010). This VTE behaviour is strongest in the early, deliberative decision-making state, and gradually subsides, becoming absent as spatial decision-making becomes automated (Redish, 2016).

To investigate the role of replay and offline activity in solving such a stochastic reinforcement learning problem, I developed and piloted a novel maze task which takes the form of a partially observable Markov decision process with stochastic rewards. In this chapter, I describe the task and characterise the behavioural performance of six rats trained on it, including sensitivity to reward and VTE behaviour. Successful performance of the task required integrating information over time, evaluating reward

probabilities to execute appropriate patterns of behaviour for maximising reward. The results show that the task successfully induced learning from stochastic rewards, which confirmed the task as a suitable behavioural tool for subsequent investigation of replay.

2.1.1. Aims of this chapter

1. To develop and pilot a novel maze task
2. To characterise the behaviour and learning performance of rats performing the task

2.2 Methods

2.2.1. Subjects

All procedures were performed in accordance with the United Kingdom Animals (Scientific Procedures) Act 1986 and European Union Directive 2010/63/EU and were reviewed by the University of Bristol Animal Welfare and Ethical Review Board. Six adult male Lister hooded rats (Charles River Laboratories, UK), weighing 260g-330g were trained on the task.

2.2.2. Materials

The three-armed radial maze consisted of a raised central platform 25cm in diameter, with three arms (60cm x 7cm) protruding from it at roughly 120° angles (fig. 2.1A). Arms were separated from the central platform by inverted-guillotine doors, which were raised to block access to the arms, and fell below the maze floor to allow access. Turning zones (10cm x 10cm) with lick ports were positioned at the end of each arm, at which 20% sucrose solution rewards were delivered. Door movements and reward delivery were operated automatically according to the rat's position, tracked using a webcam mounted above the maze, using custom MATLAB (The MathWorks) code which used changes in image brightness to identify the position of the rat on the maze in real time.

2.2.3. Behavioural protocol

Rats were trained during the light part of a 12:12 light/dark cycle, in a dimly-lit room, following at least three days of habituation to the recording room and maze-operation sounds. Each rat performed 22 training sessions lasting 1 hour each, between 5 and 7 days per week.

Trials began when a rat entered, or was placed by the experimenter on, the central platform with all doors closed. Doors opened following a 5-second delay period. When the rat reached the lick port, reward was probabilistically delivered or withheld, and doors to the other two arms were closed; the third door was closed when the rat re-entered the central platform to begin a new trial.

Each arm was assigned as either “high probability”, “mid probability” or “low probability”, which determined the protocol for reward delivery. These assignments remained fixed throughout training for each rat, but

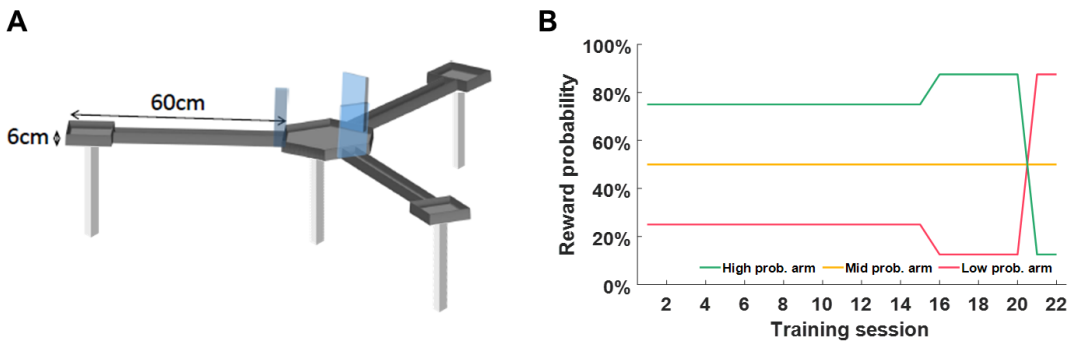

Figure 2.1. Probabilistic maze task. A. Illustration of the maze used to train animals. Lick ports located at the end of each arm delivered reward with either high, medium or low probabilities. B. Reward probabilities for each session; these applied when rats alternated between arms, but reward probability was zero for a revisit to an arm visited on the previous trial.

were counter-balanced between rats. For the first 15 training sessions, the high-probability arm delivered a reward on 6 out of 8 (75%) legitimate entries to the arm, the mid-probability arm on 4 out of 8 (50%), and the low-probability arm on 2 out of 8 (25%; fig. 2.1B). A legitimate entry was one in which a different arm had been entered on the previous trial; entering the same arm twice in a row was incorrect and did not result in a reward delivery. For sessions 16-20, the reward probabilities for the high- and low-probability arms were amplified: reward was delivered on 7 out of 8 (87.5%) and 1 out of 8 (12.5%) legitimate entries respectively. For sessions 21-22 the reward probabilities for the high- and low-probability arms were switched, such that the (formerly) high- and low- probability arms delivered reward on 1 out of 8 (12.5%) and 7 out of 8 (87.5%) of legitimate entries respectively (fig. 2.1B). Changing reward probabilities had three benefits: first, it added to the uncertainty in the task and ensured variable reward-prediction errors. Second, it allowed rats' knowledge of reward probabilities to be inferred from changes in behaviour in response to changes in reward probabilities. Third, it allowed reward probabilities to be piloted and calibrated to produce suitable behaviour for future experiments involving neural recordings (Chapter 3), e.g. an appropriate number of entries to each arm.

2.2.4. Data analysis

Vicarious trial-and-error analysis

The position of the rat was tracked using a ceiling-mounted camera at a rate of 10 frames per second, and interpolated using the MATLAB (The MathWorks) function `spline` to obtain an estimate of the position for every 10ms. For each session, k-means clustering was performed on the position data when the rat was on the central platform, which revealed three clusters adjacent to the doors, reflecting the rats' tendency to

explore the closed doors during the delay period. Each 10ms time bin was coded for the cluster d it belonged to, and the degree of VTE for trial t was calculated as:

$$V_t = 1 - \frac{n_{d_1}}{n_{d_2} + n_{d_3}}$$

where n_{d_1} is the number of time bins spent in the cluster adjacent to the door of the to-be-chosen arm d_1, and n_{d_2} and n_{d_3} are the number of time bins spent in the clusters adjacent to the other two arms, d_2 and d_3.

Probability-matching analysis

The frequency of reward at each arm was averaged over a moving window of 20 trials using LOWESS smoothing (MATLAB function `smooth`), and compared to the frequency of arm choices over the same window. The degree of probability-matching d for each session s was calculated as:

$$d_s = 1 - \sum_{a=1}^{3} \sqrt{\left(f_{a,s} - \frac{r_{a,s}}{\sum_{a=1}^{3} r_{a,s}}\right)^2} \times \frac{1}{3}$$

where $f_{a,s}$ is the frequency of choosing arm a in session s, and $r_{a,s}$ is the average reward received at rm a in session s. This gave a mean root square error between reward probability and arm choice, which was subtracted from 1 to give a measure of probability-matching where 1 is perfect probability-matching. This method is adapted from Gaissmaier and Schooler (2008).

2.3 Results

2.3.1. Learning performance

Over 22 sessions, animals learned to distinguish between the high-, mid- and low-probability arms in their frequency of visits to each arm, indicating successful learning of the reward probabilities. Rats performed 45.1 ± 2.5 trials per session (fig. 2.2A), eventually showing a significant preference for the high-probability arm and against the low-probability arm, evident by session 6 and stable by session 10. On average they distinguished between all arms on 13 out of 22 sessions (fig. 2.2C; χ^2 test, Bonferroni-corrected), visiting the arms which delivered a higher probability of reward more often, primarily in later sessions. The six rats varied in their degree of discrimination between the arms (fig. 2.2E), which may be accounted for by the orientation of the maze in the room; for example, animals may have shown a confounding preference for the arm which was closest to the door of the recording room, an effect which was overcome by rotating the arm probabilities between animals. (For rats H and K, the mid-probability arm was closest to the door; for rats I and L, the high-probability arm was closest to the door; and for rats J and M, the low-probability arm was closest to the door.)

To quantify performance on the task, each trial was coded as optimal or suboptimal according to the animal's choice of arm given the arm most recently visited. Because no reward was given for re-entering the same arm visited on the previous trial, the optimal action choice following a visit to the mid- or low-probability arm was to visit the high-probability arm; the optimal action following the high-probability arm was the mid-probability arm. The natural tendency to alternate between arms (Lalonde, 2002) was apparent from the first session (fig. 2.2B, χ^2 tests) and increased quickly to a near-100% rate of alternating. Over sessions, animals increased the proportion of trials on which they behaved optimally, achieving performance significantly above chance from session 3 onwards (fig. 2.2D, 46 trials optimal out of 106, $p = 0.02$, binomial test, Bonferroni-corrected).

2.3.2. Action latency and vicarious trial-and-error

In general, the time taken to reach the reward zone from the central platform decreased with learning, which was driven by both a decrease in time taken to enter the chosen arm (decision latency) and an increase in running speed during the reward zone approach (approach time; fig. 2.3).

Corresponding to the decrease in time to reward location, there was a tendency for rats to rear and explore the door closing off the arm they intended to enter during the 5-second delay period, roughly facing the

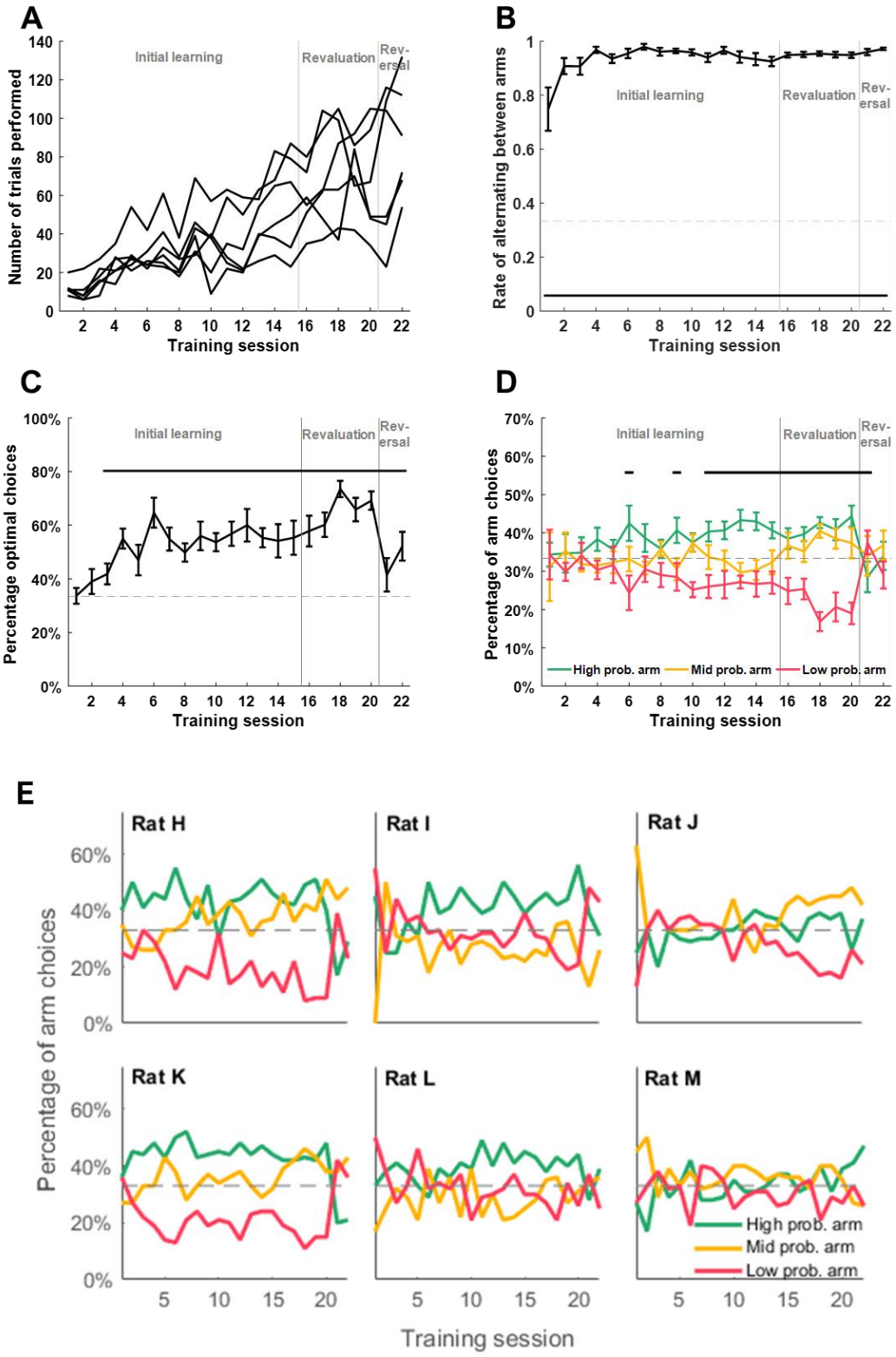

Figure 2.2. Learning performance on the maze task. **A.** Number of trials completed in each one-hour session; each plot represents one rat. **B.** Rate of alternating between arms, i.e. not revisiting the arm visited on the previous trial. **C.** Average frequency of entry to each arm. **D.** Mean proportion of trials on which the optimal arm was chosen, according to highest probability of reward. **E.** Frequency of entry to each arm over all sessions, shown separately for each rat. Black bars indicate significance, calculated using χ^2 tests. Dashed lines represent chance level (33.3%). Error bars represent standard error of the mean (s.e.m.).

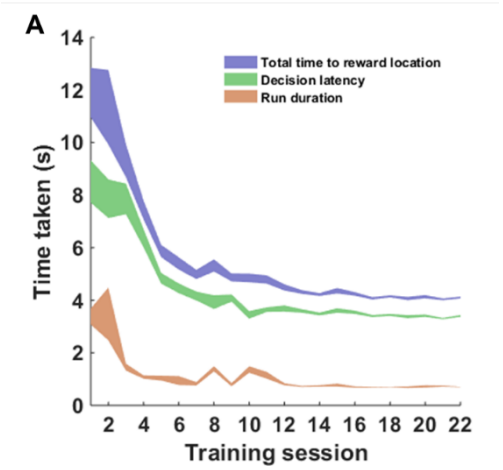

Figure 2.3. Action latency. Time from doors opening (ending the delay period) to arrival at the reward location. Total time is the sum of decision latency and run duration. Shaded areas indicate s.e.m.

direction of the reward zone (fig. 2.4A). On some trials, rats explored two or three closed doors before choosing one (fig. 2.4B), a behaviour known as vicarious trial-and-error (Muenzinger, 1931; Tolman, 1948), which has been proposed to reflect decision-making as the rat anticipates the outcome of each possible action in turn (Redish, 2016; Schmidt et al., 2013; Tolman, 1948). During VTE, hippocampal place cells show anticipatory spatial coding of possible future trajectories (Johnson & Redish, 2007), and interactions between hippocampus and other brain areas associated with decision-making show increased interaction (van der Meer and Redish, 2009; Spellman et al., 2015; Stott & Redish, 2014).

Rats' movement during the delay period at the start of each trial was coded to measure VTE (fraction of time spent at the not-to-be-chosen arms; see Methods). In line with previous reports, VTE was highest during the early stage of learning and decreased with experience (Johnson & Redish 2007; fig. 2.4C), suggesting that this behaviour might relate to deliberation, indecision and uncertainty. In later sessions, when behaviour on average was more optimal (fig. 2.2C) and less prone to VTE (fig. 2.4C), a difference emerged between the amount of VTE on optimal compared to non-optimal trials (fig. 2.4D). Specifically, on 7 out of 22 trials, VTE was greater on non-optimal than optimal trials (paired t-tests, $p < 0.05$, Bonferroni-corrected), which might reflect a decision to explore rather than exploit, requiring the kind of deliberative processing which is proposed to occur during VTE (Redish, 2016). In addition, during the first reversal-learning session (session 21) the opposite effect was true: greater VTE on optimal than non-optimal trials, when the proscribed definitions of optimal and non-optimal were changing.

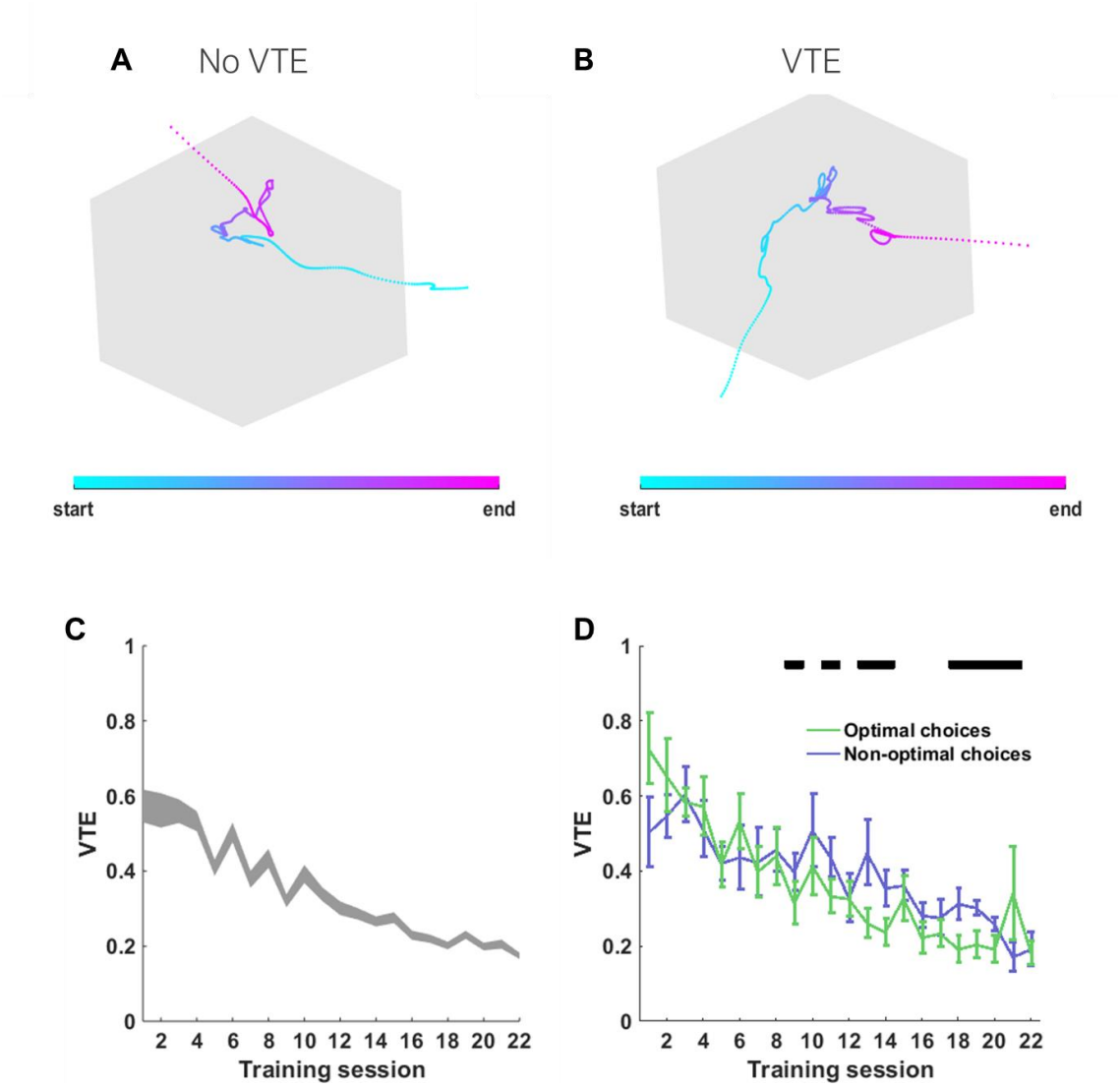

Figure 2.4. Vicarious trial-and-error. A-B. Example plots of a rat's location on the central platform (outlined in grey) from time of arrival to time of exit. A. Trajectory is mostly in one direction, with some movement corresponding to exploring around the closed door during delay period. B. Trajectory is less linear, corresponding to heading first towards one door and then towards another. C. Average VTE on all trials over learning; shaded area indicates s.e.m. D. Average VTE prior to optimal choices (i.e. entry to the arm that maximises reward) compared to non-optimal choices. Black bars indicate significant differences; error bars indicate s.e.m.

2.3.3. Exploration, exploitation and behavioural strategies

Probability-matching

Over hundreds of trials, rats' performance notably tended not towards a pattern of optimal behaviour for maximising reward, but reached an asymptote roughly in proportion to the probability of reward at each arm (fig. 2.2D). Optimal behaviour as it is proscribed here (alternating between the high- and mid-probability

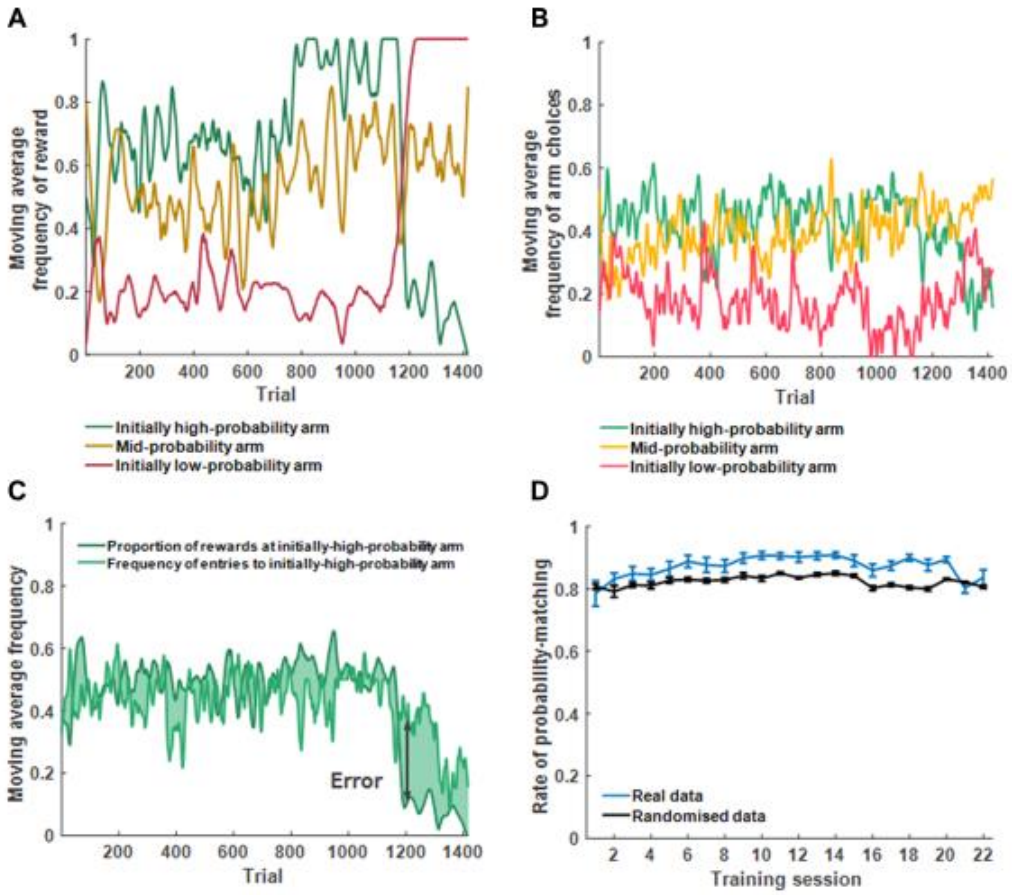

Figure 2.5. Probability-matching. A. Example average frequency of reward at each arm, over all learning trials for one rat. B. Example average frequency of choosing each arm, for the same rat. C. Calculation of the error used to analyse probability-matching: reward frequency and arm entry frequency are the same as in A and B respectively. D. Degree of probability-matching for all rats, compared to probability-matching calculated from shuffled data where actions are random.

arms to maximise reward) is contrived because it ignores the fact that there is an exploration-exploitation trade-off: an agent that solely exploits (also known as maximising) will be slow to adapt to a changing environment or may never notice that the environment has changed, while an agent that solely explores will never improve its performance (Cohen et al., 2007; Cook et al., 2013; Daw et al., 2006; Koehler & James, 2014; Vulkan, 2000). It was anticipated, therefore, that there would be some degree of exploration as well as exploitation persisting throughout training, such that rats did not reach 100% optimal behaviour. However, probability-matching is not the only possibly consequence of an exploration-exploitation trade-off (Cohen et al., 2007), so it is worth consideration.

The suboptimal pattern of probability-matching is a robust finding in tasks which involve statistical learning, particularly in human participants (e.g. Gaissmaier & Schooler, 2008; Hake & Hyman, 1953; Koehler and James, 2014; Vulkan, 2000), but also in fish (Behrend & Bitterman 1961), birds (Bullock & Bitterman 1962; Graf et al.,1964), and turtles (Kirk & Bitterman, 1965), among other species. However, there are

inconsistent results on whether rats tend to exhibit probability-matching or maximisation when presented with probability-learning tasks (Bitterman et al., 1958; Hume & Irwin, 1974; Weinstock et al., 1965) . These divergent results may be dependent on the amount of training, as probability-matching decreases with experience (Myers et al., 1963), and additionally or alternatively the tendency to probability-match early in a session and gradually maximise later in the session (Shimp, 1970).

Probability-matching has been suggested to reflect an explicit strategy of re-sampling choices that are more uncertain or which have not been sampled recently (Cohen et al., 2007). Alternatively, it has been suggested that probability-matching may result from integrating reward information from a small sample of recent trials, which are necessarily unrepresentative of the environment because of its stochasticity (Feher da Silva et al., 2017; Otto et al., 2011; Plonsky et al., 2015), or some other elaboration of the win-stay, lose-shift strategy (Worthy & Todd Maddox 2014)

Here, the degree of probability-matching in this task was quantified as the mean square error between the rate of reward at each arm (fig. 2.5A) and the rate of choosing the corresponding arm (fig. 2.5B; see Methods). Compared to random simulated data based on choosing actions randomly, real behavioural data showed that rats matched their behaviour more than predicted by chance (fig. 2.5D). This measure is very approximate, however, due to the averaging over many successive trials; the influence of win-stay behaviour on the tendency to probability-match is considered next.

State-action values

To perform this task appropriately, rats should take into account not only the average reward at the arm they choose next (the action value), but the average reward at this arm given the arm visited on the previous trial (the state-action value). This difference is essential to discriminating between many different reinforcement learning algorithms (Chapter 4), so to assess the behavioural strategy rats use, the association between outcome on one trial and action choice on a subsequent trial was investigated.

The high rate of alternating between arms (fig. 2.2B) confirms that action choices were not dominated by a simple win-stay strategy of returning to the same arm after receiving a reward. When a given state-action pair was rewarded, rats chose the same action on 1789 out of 3034 (56.8%) of subsequent trials in the same state, in the same session (fig. 2.6A). This was highly significantly different from choosing the same action by chance, i.e. 33% (binomial test, $p = 2.5 \times 10^{-183}$), and also significantly different from choosing the same action according to a simple alternation rule of not revisiting the same arm but choosing between the other two arms at random, i.e. 50% (binomial test, $p = 7.5 \times 10^{-24}$), indicating some influence of a win-stay strategy.

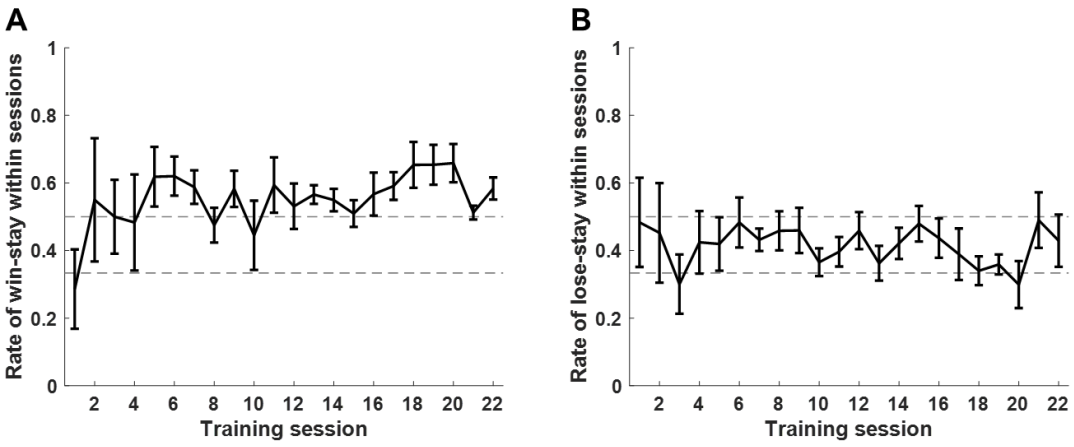

Figure 2.6. Influence of single-trial reward outcome on state-action choices. A. Rate of choosing the same action in the same state after receiving a reward (win-stay). B. Rate of choosing the same action in the same state after receiving no-reward (lose-stay). Error bars represent s.e.m; dashed lines represent two interpretations of the level of chance: 33% (according to a rule of random action selection) and 50% (according to a rule of not repeating the previous trial's action).

When a given state-action pair was unrewarded, rats chose the same action on 995 out of 2403 (40.9%) of subsequent trials in the same state in the same session (fig. 2.6B): a higher rate than chance (33%, binomial test, $p = 2.7 \times 10^{-17}$) but lower than predicted by an alternation rule (50%, binomial test, $p = 1.5 \times 10^{-126}$) and lower than the rate of win-stay (Fisher's exact test, $p = 6.4 \times 10^{-38}$). Rats were therefore more likely to choose the same action in the same state following a reward than no-reward, but this did not account for all of the variability in action choices. Further computational consideration of how reward on successive trials influences decision-making is given in Chapters 4 and 5.

2.4. Discussion

Over 22 sessions, a cohort of six rats learned to navigate for food rewards, successfully distinguishing between reward probabilities at three arms on a maze, and adapting behavioural choices to changing reward probabilities. As a pilot of a new behavioural task, these results demonstrate its suitability for investigating spatial decision-making, statistical learning, flexible behaviour, and the influence of reward on learning.

Evidence of a deliberative process in action selection was apparent from the presence of VTE, in which rats spent time exploring a closed door other than the one which lead to the arm they would eventually traverse, apparently considering and evaluating one possible choice after another. The amount of VTE, as well as the time taken to reach the reward location, decreased throughout the sessions, consistent with a gradual shift from deliberative, hippocampus-dependent action selection to automated action selection (Redish, 2016). However, a degree of VTE persisted throughout learning, especially on trials where rats eventually selected a suboptimal arm, which hints at a degree of deliberation involved in evaluating the exploration-exploitation trade-off.

Rats did not reach a point of reward maximisation, where the optimal arm choice was made consistently on every trial. Although it cannot be ruled out that they may have reached this point given enough training, the tendency to converge at a suboptimal behavioural policy is consistent with many other reports of probability-matching amongst a diverse range of species faced with probabilistic rewards. As in those studies, rats here tended to match the frequency of their entry to each arm with the frequency of obtaining rewards. They were not overly sensitive to the presence or absence of reward on an individual trial, evident from the fairly middling rates of win-stay and lose-stay behaviour, which further suggests that they were able to integrate reward information over many trials to achieve advantageous (although not optimal) behaviour on the task.

A wide range of brain areas have been implicated in the exploration-exploitation problem, including prefrontal cortex (Badre et al., 2012; Frank et al., 2009), anterior cingulate cortex (Cohen et al., 2007; Shenhav et al., 2013), locus coeruleus (Aston-Jones et al., 1994; Sales et al., 2019; Usher et al., 1999), and the basal ganglia, particularly the striatum (Beeler et al., 2010; Frank et al., 2009; Franklin & Frank, 2015; Humphries et al., 2012). The VTE behaviour which was evident at choice points on the maze recruits a similar network of brain areas which are believed to evaluate possible actions and outcomes, with the additional involvement of the hippocampus, which encodes current, future and past locations and is necessary for spatial cognition. The nucleus accumbens plays a privileged role in these processes, as the target on which many of these inputs converge to influence motor output (Floresco, 2015; Mogenson et al., 1980). It is involved in instrumental conditioning generally (Cardinal & Cheung, 2005; Cheng & Feenstra, 2006; Corbit et al., 2001; Hernandez et al., 2002; Kelley et al., 1997), has been found to convey reward-related information prior to

the point of decision-making during VTE (van der Meer et al., 2010; Stott & Redish, 2014), and has particular involvement in tasks involving uncertainty or stochastic reward (Cardinal & Howes, 2005; Dalton et al., 2014; Delgado et al., 2005; Rodriguez et al., 2006; Stopper & Floresco, 2011; Wilkinson et al., 2014). As a result, this task is appropriate for probing the involvement of the hippocampus and accumbens in learning from reward to choose future actions.

The novel task described here places a number of cognitive demands on the animal to obtain rewards. At the start of a trial, the animal must locate itself in the environment, recall where the reward locations are in the environment, estimate the probability of receiving a reward at each reward location, assess the risk-reward ratio associated with visiting each reward location, choose one, and run towards it. It was hypothesised, therefore, that activity in dorsal CA1 and the nucleus accumbens might underlie learning and performance of the task and show signatures of spatial and value learning that are updated with experience. Chapter 3 describes how we tested the involvement of these brain regions by making single-unit recordings from both areas simultaneously in one rat performing the task over multiple days.

Chapter 3: Neural dynamics of hippocampus and nucleus accumbens

3.1. Introduction

Having validated behaviour of the novel task as described in Chapter 2, simultaneous chronic recordings were made from the hippocampus (specifically dorsal CA1) and the nucleus accumbens (the ventral part of the striatum) from one rat performing the same task, to characterise the activity of these brain regions during behaviour. This chapter begins with an overview of the neurophysiological properties of the two brain regions, the characteristics of their single-unit activity and local field potential, and their anatomical and functional connectivity. Subsequently, the single-unit and LFP correlates of behaviour for this task are described, establishing these two brain regions as good candidates for the neural implementation of probabilistic reinforcement learning online and offline.

3.1.1 Neurophysiology of CA1

In mammals, CA1 of the hippocampus comprises mostly excitatory pyramidal cells, with a simple, laminar structure and highly spatially structured inputs which target distinct parts of the pyramidal cells' complex, branching morphology (Cembrowski & Spruston, 2019). Inputs arrive primarily from the CA2 and CA3 subregions of the hippocampus, which are involved in memory formation (Dudek, 2016; Kesner, 2007), the entorhinal cortex, which carries spatial information (Fyhn et al., 2004), and the nucleus reuniens of the

thalamus, which transmits information from other brain areas. The convergence of these inputs onto CA1 allows it to engage in spatial and mnemonic functions, but extensive research spanning decades has implicated it in a vast range of cognitive functions.

Classically, CA1 is characterised by place cells: pyramidal cells which exhibit sharp tuning curves selective for the animal's location in space. When the animal is in the cell's preferred place field, an area of typically a few centimetres, firing rate increases sharply. A population of place cells with overlapping place fields can therefore encode the entire environment, with the number of cells recruited to the cognitive map and the number of place fields per cell increasing as the environment grows in size (Rich et al., 2014). CA1 place cell spatial representations are both robust and flexible, forming in as little as 5-6 minutes upon exposure to a novel environment (Frank et al., 2004), but showing long-term plasticity to update as the environment changes (Rich et al., 2014; Lever et al., 2002; Muller & Kubie, 1987; Hollup et al., 2001; Dupret et al., 2010). CA1 receives input indirectly from sensory cortex (Jeffery, 2007), depending primarily on visual and olfactory input to maintain stable place cell representation (Save et al., 2000) as well as vestibular information about the animal's movement. However, the place representation quickly becomes modality-independent, performing well even in darkness after learning the environment from visual cues (McNaughton et al., 1989) and in the absence of vestibular information (Chen et al., 2013).

Despite these early findings, the purely spatial coding of place cells has been called into question. Place cell firing is also modulated by other spatial features including landmarks and borders (Gothard et al., 1996), as well as egocentric factors like head direction and running speed (Leutgeb et al., 2000; Dayawansa et al., 2006), in addition to non-spatial characteristics including the passage of time (MacDonald et al., 2011; Eichenbaum, 2014) and possibly more abstract states and structural regularities (Schapiro et al., 2016). Place cells are also highly sensitive to the animal's behavioural state, flexibly engaging or disengaging from external stimuli to switch between encoding current location and past or anticipated future locations (Yu et al., 2017; Kay & Frank, 2019). This means that downstream targets can read out a great deal of information about the animal's active state from the population activity of CA1 cells.

Unlike the accumbens, the CA1 region of the hippocampus is characterised by few recurrent connections between principal cells, but strong connections between excitatory pyramidal cells and inhibitory interneurons, resulting in a tight balance of excitation and inhibition which promotes the spontaneous generation of highly synchronous firing (Spruston & McBain, 2007). This synchronicity is apparent in the local field potential (LFP), the aggregated electrical activity of thousands of nearby neurons, as oscillations. As a result, CA1 exhibits oscillations which are strongly associated with an animal's behavioural state. The theta rhythm (6-8 Hz) dominates during active wakefulness and correlates with running speed, but also dominates during rapid eye-movement (REM) sleep. The gamma rhythm (30-100 Hz) is observed concurrently and often nested within the theta rhythm during wake, but is less closely associated with

behaviour and recruits only a small subset of hippocampal cells at a time. Sharp-wave ripples (120-250 Hz) occur as bursts lasting on the order of 100 ms during rest and especially during slow-wave sleep.

The ability of CA1 to carry such subtle and complex information is dependent on precise spike-timing, which is coordinated by these LFP oscillations to which CA1 cells are phase-locked. During running, place cells are strongly modulated by theta oscillations at roughly 6-8 Hz, and show phase precession (firing gradually earlier and earlier in the theta cycle) as the animal progresses through the place field, which allows the population to simultaneously encode current location, reflected in the firing rate of the highest-firing neuron, and immediate past and future locations, depending on the phase precession of other neurons. Nested within theta cycles are higher-frequency gamma cycles (Lopes-dos-Santos et al., 2018; Colgin, 2015), to which active place cells are also locked. The frequency of theta-nested gamma appears to change depending on cognitive demands, to allow different information flow throughout the hippocampus: high gamma power at around 80 Hz is more prevalent during learning while low to mid gamma power at around 35-55 Hz dominates during retrieval (Lopes-dos-Santos et al., 2018; Fernandez-Ruiz et al., 2017). At the higher frequency, CA1 LFP is coherent with LFP in the medial entorhinal cortex

Thus, over multiple temporal scales, CA1 can encode an incredibly complex tapestry of information about the animal's current state, based on highly structured spiking patterns. This information is communicated to nearby neurons, other hippocampal subregions, and brain areas beyond the hippocampal formation, including the striatum.

3.1.2 Neurophysiology of nucleus accumbens

The striatum has a very different electrophysiological properties. It is structurally homogenous, comprising (in rodents) 90-95% GABAergic medium spiny neurons whose inter-connections are sparse, weak and mostly non-reciprocal. Aside from a small fraction of cholinergic interneurons, internal connectivity is almost entirely inhibitory, with excitatory inputs arriving from a range of cortical, thalamic and limbic afferents, including hippocampus (Voorn et al., 2004). As a result, its principal cells are bistable (O'Donnell & Grace, 1995), firing at lower baseline rates than CA1 and switching irregularly to a high-firing state when stimulated by excitatory inputs, an effect which is modulated by strong dopaminergic inputs (Humphries et al, 2009). When strong or convergent excitatory input excites a large number of striatal cells at once, the lateral inhibition between them counteracts it with a broad dampening of activity; this raises the threshold at which afferent stimulation can excite striatal cells, producing a filtering effect. This results in very complex and chaotic dynamics, as input from one brain area can be gated by convergent inputs from another brain area (O'Donnell & Grace, 1995; Ding et al., 2010), by interneuron activity (Exley & Cragg, 2008; Goldberg & Reynolds, 2011), or by neuromodulation which alters the excitability of pre- or post-

synaptic cells (Yang & Mogenson, 1986; O'Donnell & Grace, 1996; Nicola et al., 2000; Han et al., 2007; Humphries et al., 2009). Intrinsically, however, firing in the striatum is irregular, asynchronous and generally devoid of any patterns (Buzsaki, 2006). The nucleus accumbens, forming most of the ventral part of the striatum, has been described as functionally distinct from the caudate and putamen subregions, but this is largely a result of different afferents, not differences in neurophysiology (Nicola et al., 2000; Voorn et al., 2004); in particular, the dense dopaminergic innervation in dorsal striatum arises primarily from substantia nigra while in ventral striatum it arises from ventral tegmental area (VTA). These characteristics make the accumbens well suited for integrating multiple streams of information, setting inputs from different brain areas in competition to produce a winner in situations of uncertainty, ambiguity, risk, or excessive options.

Throughout the striatum, MSNs are typically classified by their expression of D1 or D2 dopamine receptors (Gerfen et al., 1990; Surmeier et al., 1996). D1 receptors have a biphasic effect on MSNs when activated, depending on prior membrane potential: when hyperpolarised in a "down" state they reduce the cell's excitability, filtering out all but the strongest convergent excitatory drive, but when depolarised in an "up" state they enhance firing (Kreitzer, 2009). D2 receptors, in contrast, have a mainly inhibitory effect. Although MSNs also express a host of other receptors – including glutamatergic, cholinergic, cannabinoid and adenosine receptors, and even additional dopamine receptors D3, D4 and D5, to varying degrees – the presence of D1 or D2 receptors is significant because the majority of MSNs express just one or the other dopamine receptor (Bertran-Gonzalez et al., 2008), roughly equally divided between the two, with corresponding differences in circuitry and morphology. Morphologically, D1 cells have more dendrites, receiving more glutamatergic input, and are therefore less excitable (Gerfen & Surmeier, 2011). The two populations receive similar inputs, but D1-expressing cells generally project to the basal ganglia's output nuclei (specifically the internal segment of the globus pallidus and the substantia nigra as well as VTA), forming a direct pathway which disinhibits a selected action. D2-expressing cells, meanwhile, send signals to the output nuclei indirectly by projecting via the ventral pallidum and external segment of the globus pallidus, ultimately producing an inhibitory effect which impedes selected actions. Activation of these receptors is therefore said to form parallel "go" and "no-go" signals, respectively. Excitatory inputs to striatum can activate ensembles of cells which encode a particular action; if this action is encoded by D1 cells the signal drives initiation of the action, whereas if it is encoded by D2 cells the signal drives inhibition of the action.

Because of the properties of D1 and D2 receptors, dopamine has divergent effects on these populations of cells. Phasic dopamine release by VTA promotes long-term potentiation (LTP) at synapses on D1 cells, which serves to reinforce connections associated with the positive reward-prediction error which elicited the dopamine transient, but promotes long-term depression (LTD) at synapses on D2 cells. D1 receptors are expressed postsynaptically, modulating the response of neurons to neurotransmitter release, but D2 and D3 receptors can be expressed both presynaptically and postsynaptically, making their influence more

complex. Presynaptic action regulates the release of neurotransmitters at the synapse and provides autoinhibiting negative feedback to control firing rates, and D2 receptors are activated by lower concentrations of dopamine, which means that dopamine can have a biphasic effect depending on the concentration (Beaulieu & Gainetdinov, 2011). D2 receptors are found not only in striatal cells (i.e. MSNs and striatal interneurons) but also presynaptically on the axon terminals of afferent cells, controlling the release of GABA and glutamate (Bamford et al., 2004). This means that phasic dopamine release by VTA can inhibit glutamate release by hippocampal inputs (Yang & Mogenson, 1984), preferentially targeting weak inputs, effectively filtering out all but the strongest convergent excitatory input.

Although this dichotomous D1/D2 classification characterises dorsal striatal medium spiny neurons with some success, it is overly simplistic and somewhat murky especially for accumbens. One challenge to the framework is that the proportion of MSNs co-expressing both D1 and D2 receptors is roughly 5% for dorsal striatum and 6% for the core part of the accumbens, but as high as 17% for the shell part surrounding the core (Bertran-Gonzalez et al., 2008). These co-expressing cells form a distinct class, with smaller cell bodies and dendritic arbours, and a behavioural significance that is unknown (Gagnon et al., 2017). D3 receptors are highly co-expressed on D1 cells, despite belonging to the same family as D2 receptors, and are especially abundant in accumbens shell (Schwartz et al., 1998). Moreover, the accumbens core-mediated indirect pathway contains a large proportion of both D1 and D2 MSNs (Kupchik et al., 2015), making the dissociation less clear-cut. Both populations in NAc can inhibit or disinhibit target cells in thalamus, irrespective of their expression of D1 or D2 cells (Kupchik et al., 2015), suggesting that this receptor expression is less important than in dorsal striatum for determining circuitry and function. Similarly, the association of D1 and D2 cells with a “go” and “no-go” signal respectively is mixed in the accumbens: inactivation of D1 receptors reduces approach towards reward and inactivation of D2 receptors reduces avoidance in a conflict task (Nguyen et al., 2018), but phasically stimulating either cell type can drive motivation to act (Soares-Cunha et al., 2016).

MSNs form connections with roughly 10% of other MSNs, providing weak lateral inhibition which at its simplest produces “winner-takes-all” competition, whereby the cell which is excited the mostly strongly inhibits its neighbours and is in turn further disinhibited. But more subtle dynamics between MSNs are also observed: when membrane potential is elevated to just around the spiking threshold, mild inhibition can briefly hyperpolarise a cell enough to delay (rather than completely suppress) spiking, adjusting spike timing to promote synchronised spiking in the population (Plenz, 2003). And because MSNs are highly polarised at rest, with a resting potential of -80 to -90 mV, GABAergic input can actually have a depolarising effect (in contrast to its usual hyperpolarising effect throughout the brain). Although GABAergic stimulation cannot cause excitation, lateral connections between MSNs can therefore facilitate excitatory inputs which arise immediately following GABAergic input when the membrane is still depolarised (Plenz, 2003).

The other 5-10% of cells in the striatum, besides medium spiny neurons, are interneurons, which comprise four classes. Three classes are GABAergic, defined by their expression of parvalbumin, somatostatin and calretinin, respectively. Parvalbumin-positive interneurons are noticeable by their high tonic firing rates, and, much like fast-spiking interneurons found elsewhere in the brain, form proximal synapses on local principal cells (i.e. MSNs) to inhibit firing as well as gap junctions with other interneurons to promote firing synchrony and entraining oscillations. Somatostatin-positive and calretinin-positive interneurons exhibit lower firing rates and burst firing, and are less well studied. All three GABAergic interneuron types receive glutamatergic and dopaminergic innervation from outside the striatum. The fourth class of interneurons are large, slow-firing cholinergic interneurons, which express D2 and D5 receptors (the latter from the same family as D1 receptors, with similar mechanisms of action), which suppress acetylcholine and enhance responsiveness to GABA. These interneurons form excitatory connections with both MSNs and parvalbumin-positive interneurons, and have a strong effect on dopaminergic terminals: not only does cholinergic activation enhance dopamine release but it also initiates it spontaneously, even without an increase in midbrain firing rates (Cachope et al., 2012; Yorgason et al., 2017). Despite making up just 1% of striatal cells, cholinergic interneurons have received more attention for their dense arborisation and the strong influence they have on behaviour: elevated firing is associated with states in which motivation is low, including satiety and depressive-like mood, and notable pauses in tonic firing appear in response to salient reward-predictive cues which prompt behaviour (Collins et al., 2019). The actions of dopamine on cholinergic interneurons shows topographical differences: in dorsal striatum cholinergic interneurons pause in response to the firing of dopamine neurons via activation of D2 receptors; in the medial shell the pause is preceeded by a burst in firing mediated by glutamate co-released with dopamine; and in the core the effect is weaker and mixed (Chuhma et al., 2014). Dopaminergic neurons corelease glutamate in accumbens (but not in dorsal striatum; Stuber et al., 2010; Tecuapetla et al., 2010). These burst-pauses may also permit dopamine-dependent plasticity to take place, which is otherwise blocked by muscarinic action (Berke, 2018). In addition to dopamine release, VTA modulates striatal cells in other ways: GABAergic input inhibits cholinergic interneurons to increase associative learning (Brown et al., 2012), while glutamatergic excitation of GABAergic interneurons drives aversive behaviour (Qi et al., 2016).

In addition to the somewhat fluid division between dorsal and ventral parts of the striatum, the accumbens part of ventral striatum is further subdivided into a central core, adjacent to the anterior commissure, and a shell region which surrounds it (Groenewegen et al., 1999; Zaborszky et al., 1985; Zahm, 2000). These compartments are distinguished immunohistochemically and by their connections: the core, in fact, bears histochemical similarities to the dorsal striatum, as well as similar size of cells, spine density, and efferents, with the shell more clearly distinct (Humphries & Prescott, 2010). As in dorsal striatum, MSNs expressing D1 or D2 receptors are homogeneously distributed throughout the core, but D2 cells are distributed unevenly in the shell (Gangarossa et al., 2013). Shell MSNs also express D3 receptors, whereas core MSNs (like dorsal striatum) show low levels of D3 expression (Gangarossa et al., 2013). Functionally, selective inactivation shows a double dissociation in their roles in reward-related behaviour and learning:

core is involved in driving approach towards reward-related stimuli, while shell is involved in learning the irrelevance of stimuli. Although most afferents project to both core and shell, topographical gradients are apparent in the density of their projections. For example, while both areas receive input from the prelimbic and agranular insular cortices, the dorsal part of both areas projects preferentially to core and ventral to shell (Voorn et al., 2004). Dorsal CA1 projects directly to both areas (Trouche et al., 2019), but ventral CA1 axons are found with greater density in the medial shell and dorsal CA1 axons in the core and lateral shell (Voorn et al., 2004). The efferent projections, likewise, show some topography, with core and shell exhibiting preferential targeting of dorsolateral and subcommissural ventral pallidum respectively, a core bias towards substiantia nigra and a shell bias towards ventral tegmental area (VTA) and amygdala (Heimer et al., 1991). Dopamine concentrations are greater in the shell, and cell density is greater, while cells in the core have greater dendritic lengths.

In addition to the dorsal/ventral and core/shell divisions, the striatum is also anatomically segregated into patches or striosomes and surrounding matrix compartments, distinguishable by immunohistochemical staining. Despite some differences in afferent and efferent connections between the two, the functional significance of this compartmentalised structure is not well understood, and particularly under-studied in accumbens (Brimblecombe & Cragg, 2016). Patches comprise about 10% of the volume of the striatum, and in accumbens they preferentially innervate dopaminergic VTA neurons. Throughout the striatum, the main source of dopamine to patches is substantia nigra pars compacta, but dopaminergic innervation of the matrix shows differences between dorsal and ventral parts of the striatum: the accumbens matrix receives less input from substantia nigra and more from VTA. The distinction between patch and matrix has been investigated in more detail for dorsal striatum than accumbens, and evidence suggests that there are important differences between the two areas: for example, evoked dopamine release in dorsal striatum is greater in matrix than patches, but greater in patches in accumbens (Salinas et al., 2016); further differences in dorsal and ventral patch-matrix organisation may also exist but are unknown. The dendritic arbours of the majority MSNs respect the patch-matrix boundary, being restricted to just one compartment, resulting in little cross-communication between compartments; DA cells tend to project to one or the other; interneurons tend to occupy the patch-matrix boundary (Brimblecombe & Cragg, 2016). Calbindin-positive interneurons are found preferentially in the matrix.

3.1.3 Anatomical and functional connectivity between CA1 and nucleus accumbens

Being relatively slow, the theta rhythm is suited for coordination of neural activity across long distances in the brain because it can tolerate conduction delays (Colgin & Moser, 2010). Theta power in CA1 therefore often coincides with theta power in other brain areas – particularly striatum, septum, prefrontal cortex and entorhinal cortex – and inter-area coherence at this frequency is often associated with behaviours that

require complex information processing such as decision-making (Jones & Wilson, 2005; Benchenane et al., 2010; Sirota et al., 2008). High theta coherence between CA1 and accumbens is observed at choice points on mazes or upon presentation of an instructive cue (Lansink et al., 2009; van der Meer & Redish, 2011; Lansink et al., 2016), suggesting a mechanism of coordination between these two regions by which spatial information is conveyed to the accumbens (Sjulson et al., 2018).

In addition to theta oscillations which may originate from the hippocampus (Lalla et al., 2017), the accumbens shows high power at other frequency ranges which are selectively coherent with other brain regions. Like CA1, accumbens exhibits oscillations in the gamma range: it is dynamically coherent with amygdala at low gamma frequencies (35-45 Hz; Popescu, Popa & Pare, 2009), piriform cortex at mid gamma frequencies (~50 Hz; Carmichael et al, 2017), and prefrontal cortex at high gamma frequencies (80-100 Hz; Berke, 2009). This may be a mechanism for switching to the influence of inputs from one brain area over others.

Accumbens entrainment to hippocampal theta-band LFP is possible because of a direct projection from the latter to the former. Both dorsal and ventral CA1 pyramidal cells project preferentially onto parvalbumin-positive (PV+) interneurons but also onto MSNs to recruit assemblies of accumbens cells during the retrieval of reward-related spatial memory (Scudder et al., 2018; Trouche et al., 2019), by increasing the firing rate of interneurons, which entrains the spiking of MSNs. As a result, accumbens cells become tightly coupled to hippocampal theta oscillations just before (van der Meer & Redish, 2011) or after (van der Meer et al., 2019) reward delivery, with an anticipatory firing pattern that shows phase precession relative to hippocampal theta (van der Meer & Redish, 2011). The accumbens cells which are entrained by hippocampal theta encode more spatial information (Sjulson et al., 2018), demonstrating a functional link. Ventral hippocampus also projects to the accumbens (Goto & O'Donnell, 2001), and these connections are also necessary for spatial reward learning (LeGates et al., 2018), but they recruit different ensembles of cells during spatial learning (Sosa et al., 2019). The distinction between accumbens ensembles associated with dorsal and ventral hippocampus respectively is particularly apparent during sharp-wave ripples: a subset of accumbens cells show changes in firing in response to ripples (Goto & O'Donnell, 2001; Pennartz et al., 2004; Lansink et al., 2008; Sjulson et al., 2018), but ripples occur asynchronously in dorsal compared to ventral hippocampus and only the former encode task-related information during a spatial learning task (Sosa et al., 2019).

CA1 input to accumbens also shows preference for anatomical regions and cell types within accumbens. In accumbens core, stimulation of ventral hippocampus excites D1 MSNs more strongly than D2 MSNs (MacAskill et al., 2014), but produces equal inhibition of D1 and D2 MSNs via activation of interneurons (Scudder et al., 2018). Phasic dopamine release augments hippocampal inputs to accumbens core by stimulating D1, and not D2, receptors, an effect which scales with the strength of the hippocampal input (Goto & Grace, 2005). Hippocampal inputs often converge with medial prefrontal cortical inputs (as well

as inputs from other brain areas) onto the same cells (French & Totterdell, 2002), and the D1-mediated effect of phasic dopamine release makes the cells more responsive to hippocampal excitation at the expense of responsiveness to medial prefrontal cortex (Goto & Grace, 2005); activation of D2 receptors by tonic dopamine has the opposite effect. Phasic activation of D1 receptors also promotes long-term potentiation at hippocampal-accumbens synapses. This has behavioural significance: infusion of a D1 antagonist into accumbens slows the hippocampus-dependent aspect of reward-driven spatial learning, but infusion of a D2 antagonist does not (Goto & Grace, 2005).

3.1.4 Cognitive function of the nucleus accumbens

The result of such rich afferent dynamics with various cortical and subcortical brain areas is that the accumbens serves as a "switchboard" (Gruber et al., 2009), flexibly integrating information carried by disparate networks across the brain to bias action selection in the motor output system. Accumbens projection cells primarily target ensembles in the ventral pallidum, subsantia nigra and VTA (Kupchik & Kalivas, 2017; Heinsbroek et al., 2016), which variously inhibit and disinhibit parts of the thalamo-basal ganglia system to influence action selection. Lesion and pharmacological or optogenetic inactivation studies have found that the accumbens is not necessary for "simple" learning, e.g. discriminating between cues or actions to obtain a food reward (Floresco et al., 2006), or discriminating between larger and smaller rewards (Ghods-Sharifi & Floresco, 2010; Salamone et al., 1994; Stopper & Floresco, 2010). But behavioural flexibility such as set-shifting (Floresco et al., 2006; Block et al., 2007; Haluk & Floresco, 2009), discriminating between many stimuli on an eight-arm radial maze (Schacter, Yang, Innis & Mogenson, 1989), and evaluating relative costs and benefits under conditions of uncertainty or delay-discounting (Cardinal et al., 2001; Ghods-Sharifi & Floresco, 2010; Stopper & Floresco, 2010; Hauber & Sommer, 2009) are impaired under inactivation of the accumbens. What is notable is that similar behavioural impairments are seen when the projections from other brain areas to the accumbens are disrupted (e.g. anterior cingulate cortex, Hauber & Sommer, 2009; prefrontal cortex, Block et al., 2007, Feja & Koch, 2014, Groman et al., 2019; hippocampus, Trouche et al., 2019; amygdala, Bercovici, 2018, Groman et al., 2019), indicating that the accumbens is crucial for transmitting spatial, motivational and contextual information conveyed by other cortical and subcortical structures to the motor output system.

In particular, the accumbens plays a privileged role in probabilistic learning tasks (Dalton et al., 2014; Groman et al., 2019). Reversal learning is unaffected by accumbens inactivation in tasks involving deterministic rewards, but impaired when rewards are probabilistic (Dalton et al., 2014). Similarly, accumbens inactivation only marginally reduces the ability to choose a large deterministic reward over a small one, but has a more disruptive effect on the ability to choose a large but risky deterministic reward over a small but certain one, when the riskier reward is more advantageous (Cardinal & Howes, 2005;

Stopper & Floresco, 2010). In humans, increased activation of the ventral striatum accompanies probabilistic learning (Koch et al., 2008; Delgado et al., 2005; Cools et al., 2002; Mell et al., 2009), and in particular the processing of probabilistic feedback (Wilkinson et al., 2014; Rodriguez et al., 2005).

Unsurprisingly, given the broad involvement of the accumbens in processing different task-dependent information, individual accumbens neurons have been found to encode a wide variety of stimuli, actions, outcomes and errors. Prior to and during action execution, accumbens firing has been reported which is selective for cue identity (Wilson & Bowman, 2005; Gmaz et al., 2018; Day et al., 2007), upcoming action (Roesch et al., 2009), prediction of reward outcome (Wilson & Bowman, 2005; van der Meer & Redish, 2009; Bissonette et al., 2013; Strait et al., 2015), locomotion (Pennartz et al., 2004; Lansink et al., 2008; Sjulson et al., 2018), and current action (Day et al., 2007). After action execution, responses to current action (Kim et al., 2009), reward outcome at a particular location or any location (Donnelly et al., 2015; Kim et al., 2009; Lansink et al., 2008; Nicola et al., 2004; Atallah et al., 2014), reward outcome on the previous trial (Kim et al., 2009), and reward-prediction error (Kim et al., 2009) have been reported, and often a combination of these factors. Timing, also, is often a particular characteristic of single-cell responses: cells not only encode a task-relevant feature, but reliably exhibit the response before, during, after or throughout an event, or at more than one timepoint, with a response that may last milliseconds or tens of seconds (Berke et al., 2009). Relatively characteristic of the accumbens is the observation of "ramping" activity of neurons: a gradual increase (or sometimes decrease) of firing rate leading up to a response or outcome (Khamassi et al., 2008; van der Meer & Redish, 2009; Donnelly et al., 2015; van der Meer et al., 2010) which may reflect anticipation or motivation. Such temporal specificity has been suggested to underlie timekeeping (Donnelly et al., 2015) or credit assignment (Gmaz et al., 2018).

Both regions, then, exhibit complex dynamics and coding properties which are essential for cognitive functions. Given the spatial coding of CA1 and the coding of reward- and uncertainty-related information in accumbens, these are good candidate areas for investigating the coordination of activity over time for solving a stochastic reinforcement learning problem. Before considering the offline activity of these brain regions in chapter 6, I characterise their function during the task in this chapter by describing the neural correlates of behaviour and interactions, to identify the neural activity that might be replayed to consolidate new information.

3.1.5 Aims of this chapter

1. To identify behavioural correlates of CA1 and accumbens activity on the probabilistic maze task
2. To verify the presence of task-related functional connectivity between CA1 and accumbens which could be the subject of subsequent replay

3.2. Methods

One rat was implanted with dual-site 64-channel silicon probes, directed at dorsal CA1 and nucleus accumbens respectively, and trained over 17 sessions on the same maze task described in chapter 2.

The electrophysiological data from just one rat are presented in this thesis, because attempts to conduct the experiment with six more subjects faced overwhelming problems. The reasons for this were, non-exclusively, extended troubleshooting involving the maze (one rat), broken silicon probes (one rat), difficulty targeting the CA1 pyramidal layer (three rats), and unstable implants which became detached from the skull (four rats).

3.2.1 Silicon probes, animals and surgery

All procedures were carried out in accordance with the United Kingdom Animals (Scientific Procedures) Act 1986 and European Union Directive 2010/63/EU and reviewed by the University of Bristol Animal Welfare and Ethical Review Board. One male Lister hooded rat (Charles River Laboratories, UK), weighing 375g at the time of surgery, was implanted with a 9mm H2 probe and a 9mm E2 probe (Cambridge NeuroTech, UK) in the hippocampus and accumbens, respectively (fig. 3.1A). Probes were mounted on aluminium blocks (7.5mm x 3.3mm x 3.0mm) and targeted at 2.1mm lateral, 4mm posterior and 2.5mm ventral to bregma (CA1) and 1.5mm lateral, 1.7mm anterior and 7mm ventral to bregma (accumbens) respectively, in the right hemisphere, based on the atlas of Paxinos and Watson (1996). Histology to verify the probe locations was not possible because the probe implant became detached before the procedure could be carried out, but histology carried out on another rat with probes using the same accumbens coordinates verified the positioning. Behavioural training began 38 days after surgery.

3.2.2 Neurophysiological recordings

Extracellular recordings were made using an Open Ephys acquisition system at a sampling rate of 30kHz, with two RHD2164 headstages, one with an integrated accelerometer. Recordings were referenced to a stainless steel screw implanted over the cerebellum. In a typical recording session, a rest period of two hours was recorded while the rat rested in its home cage, followed by one hour of maze training, and second two-hour rest period. A red LED was attached to the implant, and the session was recorded by a ceiling-mounted webcam which allowed the rat's movement to be tracked using custom MATLAB (TheMathWorks) code. Electrophysiological recordings and position tracking were synchronised post-hoc using a second LED which blinked at random intervals.

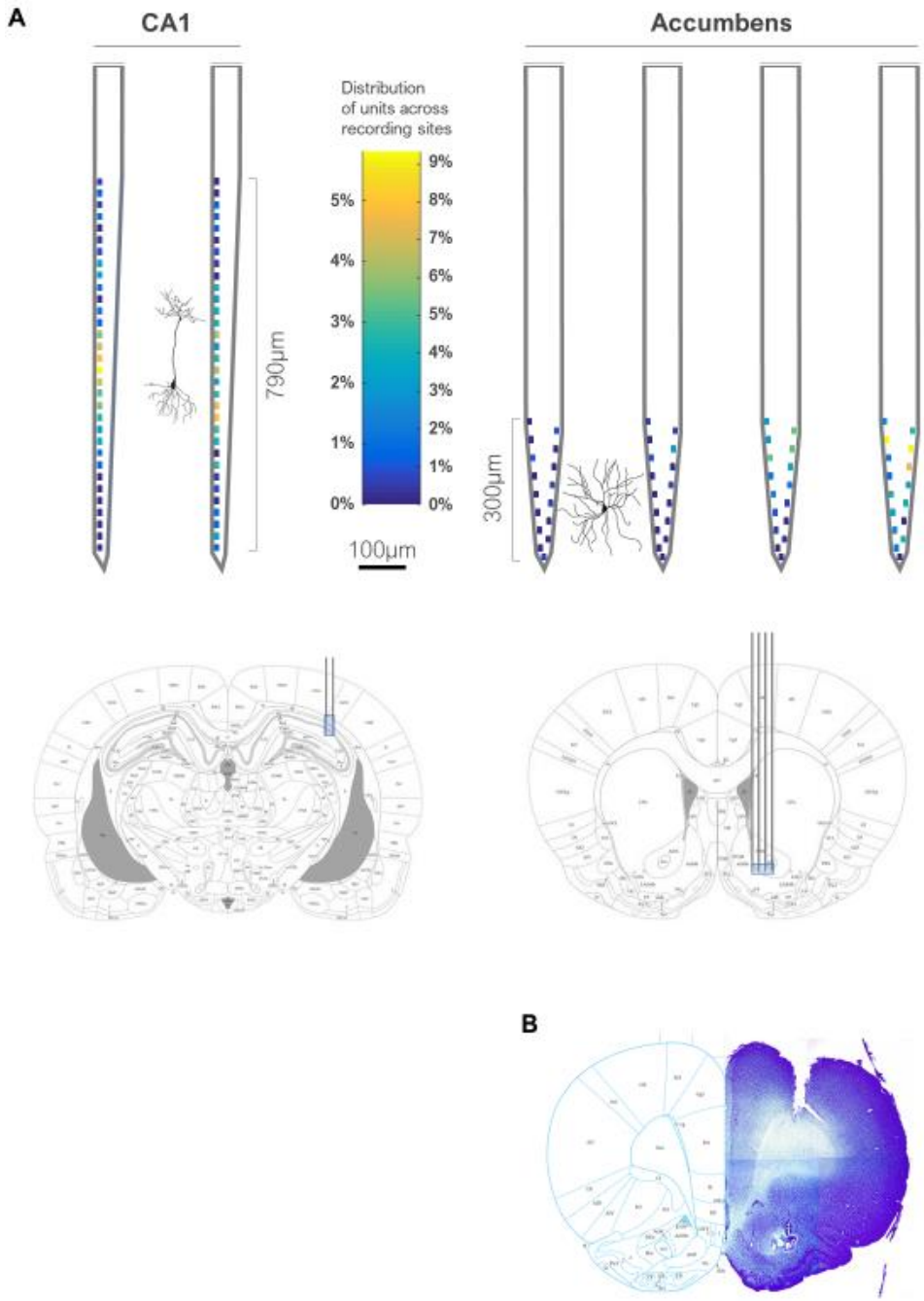

Figure 3.1. Silicon probe recordings. **A.** Top: schematic of recording sites on CA1- and accumbens-targeting probes. Colours indicate the percentage of all recorded cells whose highest amplitude was found at each recording site. Hippocampal pyramidal cell (left) and striatal medium spiny neuron (right) are shown with cell body sizes to scale. Bottom: probes shown to scale in their target locations (CA1: 2.1mm ML, -4.0mm AP, 2.5mm VD. NAc: 1.5 ML, +1.7 AP, 7.0 VD). Shaded box indicates area where recording sites were. Images adapted from Paxinos & Watson (1996). **B.** Histology from another rat with an identical probe targeted at the same accumbens coordinates; electrolytic lesions indicate the location of recording sites.

3.2.3 Behavioural training

The rat was habituated to handling and the recording room over several weeks, but otherwise there was no pre-training prior to the first recording session. Following a week's recovery post-surgery, the rat was food-restricted to no less than 85% of its body weight, and behavioural training began one month later. 17 sessions were completed in total; one session (session 9) was excluded from analysis because the position-tracking data could not be recovered, and sessions 1 and 2 were also discarded from most analyses because the rat's behaviour was erratic. Except where otherwise stated, analyses were therefore performed on the remaining 14 sessions.

Informed by the behavioural responses to changing reward probabilities presented in Chapter 2, the initial reward probabilities on the three maze arms were 87.5%, 50% and 12.% respectively. After 12 sessions, the reward probabilities at the high- and low-probability arms were switched to induce reversal learning. All other aspects of the behavioural protocol were the same as those outlined in Chapter 2.

3.2.4 Spike-sorting

Raw data were automatically spike-sorted using Kilosort software (Pachitariu et al., 2016) and manually curated using Phy (Hunter et al., 2015). In brief, raw data were common-average referenced, high-pass filtered and whitened to remove correlated noise, before prototypical spikes were detected whenever the amplitude exceeded a given threshold. Detection and clustering of dimensionality-reduced spike waveforms were then optimised iteratively using a template-matching procedure. In the manual curation step, clusters were merged, accepted or rejected as noise by visual inspection, according to their inter-spike interval histograms, amplitude, and spike waveform. Finally, clusters were restricted to those with an isolation distance of >15 (Schmitzer-Torbert et al., 2005).

3.2.5 Data analysis

Local field potential. Local field potential (LFP) was taken from one channel on each probe. For the accumbens probe, this was the channel with the strongest gamma power, averaged over one session; for the CA1 probe, this was the channel with the greatest ripple amplitude. Bandpass filtering was done using the `eegfilt` function from the EEGLAB MATLAB toolbox. LFP coherence was calculated using the `cohgramc` function from the Chronux toolbox, with a time window of 500ms, step size of 50ms, and tapers [3 5].

Discrimination of principal cells from interneurons. Units in accumbens with mean firing rate > 2 Hz and spike half-amplitude width < 0.265 ms across the whole recording session were classified as putative interneurons, consistent with previous classifications of striatal interneurons (Berke et al., 2004; Berke et al., 2008); all other cells were classified as putative medium spiny neurons. Based on previously described characteristics of CA1 interneurons and pyramidal cells (Csicsvari et al., 1999; Harris et al., 2001; Quirk & Wilson, 1999; Rank, 1973), as well as apparent clusters in the data, units in CA1 with mean firing rate > 2 Hz, spike half-amplitude width < 0.25 ms and complex-spike fraction < 0.03 across the whole recording session, as well as any unit with a firing rate > 20 Hz, were classified as putative interneurons; all other cells were classified as putative pyramidal cells.

Reward-related firing. Following Lansink et al. (2008), a cell's firing in the 2-second period centred on arrival at reward location was binned in 250ms bins, and compared to firing in three control bins on the same task (in this case, the 750ms window from entry to the central platform). A cell was considered to

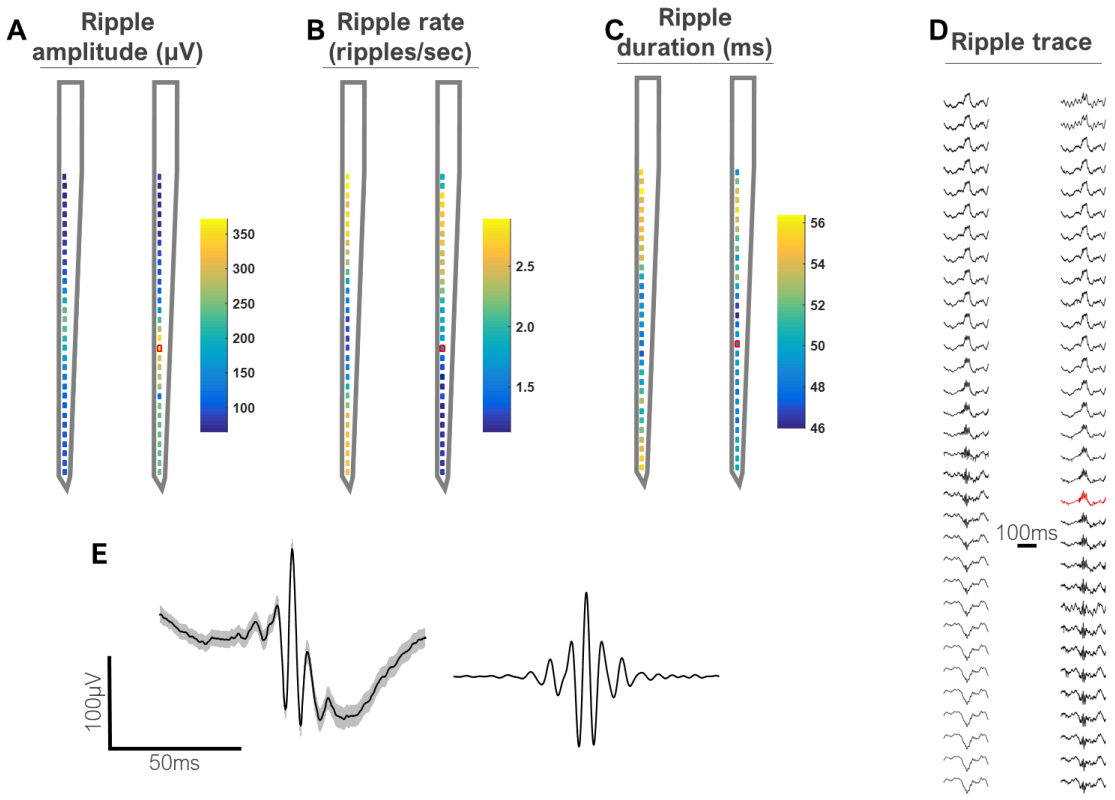

Figure 3.2. Ripple detection. A-C. Properties of ripples detected using the LFP recorded at each recording site respectively. Red border indicates the site which was eventually chosen for ripple analyses; this was the site which gave the highest ripple amplitude. D. Example trace of one ripple as recorded on every channel; red trace indicates recording site used for ripple analyses. E. Average ripple trace of all ripples recorded during one 60-minute TASK period, raw (left) and filtered at 120-250 Hz (right).

show reward-related firing if firing in at least one of the reward bins was significantly different from all three control bins, assessed using a rank-sum test with alpha value $p = 0.01$.

Ripple detection. Discrete ripple events were detected using custom MATLAB code. Raw LFP from one probe recording site was filtered at the 120-250 Hz range, and periods when the ripple power exceeded 2 standard deviations above the mean (TASK period) or 3 standard deviations above the mean (PRE and POST-task rest) were extracted. Ripple events with power exceeding 25 standard deviations were rejected as noise. Events with a duration of 10-500ms, amplitude of 30-1000μV and amplitude were preserved, and those separated by less than 30ms were merged. To determine the best CA1 probe recording site to use for ripple detection, ripple detection was initially run on all sites. The site which elicited the greatest average ripple amplitude was ultimately selected for analysis (fig. 3.2). Ripples during TASK were further restricted to those which occurred outside theta epochs: specifically, ripples were excluded if they occurred within 500ms of a theta cycle peak which exceeded 0.5 standard deviations above the mean amplitude of theta peaks during TASK.

Ripple-modulation of single-cell firing rates. To determine whether a cell's firing rate was modulated by ripples, all ripples from either the PRE, TASK or POST periods were extracted, and the firing rate within 250ms of every ripple was binned in 10ms bins. This was compared to the cell's baseline firing rate in the 200ms before and after each ripple window. A z-test was conducted on each 10ms bin to determine

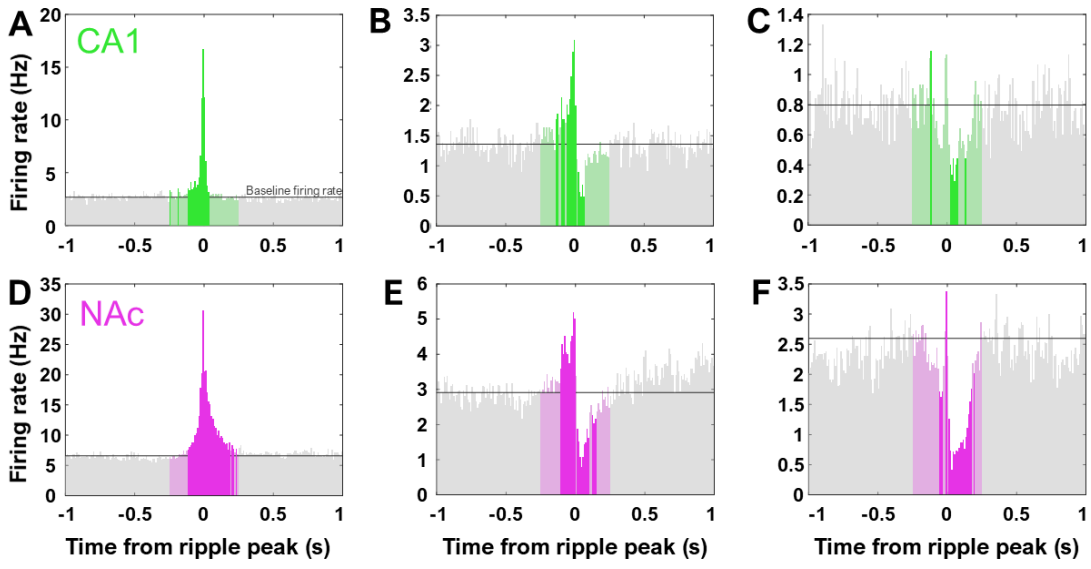

Figure 3.3. Ripple-modulation of firing rates. Ripple-triggered firing rates of 6 example ripple-modulated cells. Bars are 10ms wide. Coloured bars indicate those within 250ms of the ripple peak which were used for significance testing. Horizontal line indicates the baseline firing rate outside ripples. Brightly coloured bars are those in which the firing rate deviated significantly from baseline. **A-C.** CA1 cells. **D-F.** Accumbens cells.

whether the firing rate within the bin was different from the concatenated baseline firing around all ripples. Cells which showed significantly increased firing in at least five 10ms bins out of 500ms, at an alpha value of $p = 10^{-15}$, were assigned as positively ripple-modulated (fig. 3.3A & 3.3D). Cells which showed significantly decreased firing in at least five bins were assigned as negatively ripple-modulated (fig. 3.3C & fig. 3.3F). A large number of cells showed both positive and negative modulation; these were treated as positively modulated for analyses (fig. 3.3B & fig. 3.3E).

Theta modulation. For each cell, a theta modulation index was calculated based on all spikes fired during the TASK period (Cacucci et al., 2004). Inter-spike intervals between all spikes were obtained, the theta trough was calculated as the mean firing rate at an inter-spike interval of 50-70ms, and the theta peak was calculated as the mean firing rate at an inter-spike interval of 100-140ms. The theta modulation index was calculated by taking the difference between these values and dividing by their sum (fig. 3.4).

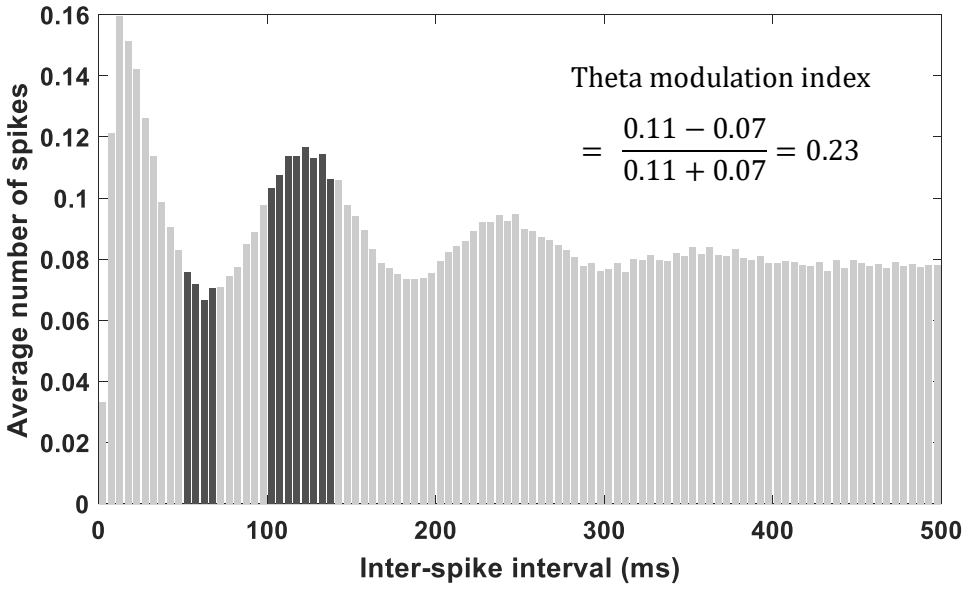

Figure 3.4. Theta modulation. Illustrative histogram of inter-spike intervals for an example putative CA1 pyramidal cell. Each bar shows the average number of spikes fired at the specified interval after a previous spike; bars are 5ms wide. Darker bars correspond to intervals of 50-70ms and 100-140ms. The cell's theta modulation index was calculated by taking the difference between the average number of spikes at these intervals and dividing by their sum, as shown (values are rounded to two significant figures).

3.3. Results

3.3.1 Behaviour

The rat was trained on the stochastic reinforcement learning maze task over 17 sessions, performing an average of 15.5 ± 1.2 trials per one-hour session (fig. 3.5A), notably less than the rats described in Chapter 2. Successful learning of the task requires alternating between the high- and low-probability arms to maximise reward; like the cohort described in Chapter 2, the rat showed a correct tendency not visit the same arm more than once in a row (fig. 3.5B), and avoidance of the low-probability arm (fig. 3.5D). As a result, it reached a similar rate of optimal performance during the initial learning stage of sessions 1-12 (fig. 3.5E), although showed little reversal learning in sessions 13-17 when reward probabilities were changed. The time taken to reach the reward location once doors blocking access were opened generally decreased over learning, but remained variable, perhaps reflecting deliberation or indecision on some trials (fig. 3.5F).

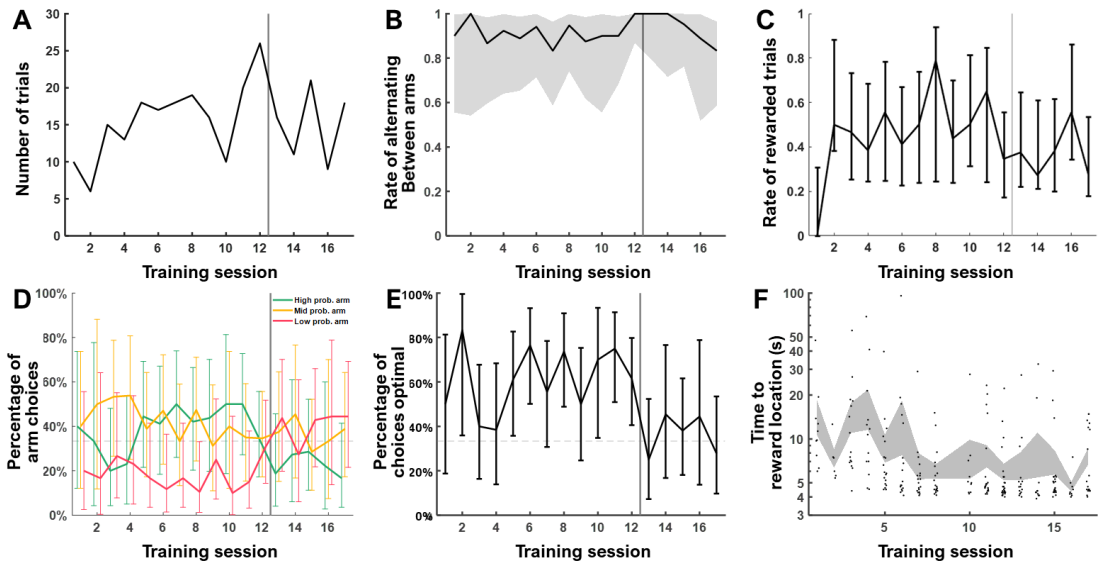

Figure 3.5. Learning performance. A. Number of trials completed per session. B. Rate of switching between arms, i.e. not visiting the same arm twice in a row. C. Proportion of trials which were rewarded. D. Proportion of trials on which the rat visited the high-, mid- and low-probability, arms respectively. E. Proportion of trials which were optimal, i.e. alternating between high- and mid-probability arms. F. Time from doors opening to arrival at the reward location for every trial. Shaded areas and error bars in B-E represent binomial confidence intervals. Shaded area in F represents s.e.m. Vertical lines indicate when

3.3.2 Single-unit activity

A total of 655 units were recorded over 17 recording sessions, after excluding those with low isolation distance: 430 from CA1 and 225 from accumbens. These were divided into principal cells and interneurons as described in the Methods. For most analyses, cells from sessions 1, 2 and 9 were excluded (see Methods for details), leaving 343 from CA1 and 169 from accumbens remaining. Cells generally showed similar characteristics to those reported in the literature (fig. 3.6).

Trials were divided into five behaviourally-relevant timepoints (fig. 3.7): the rat's entry to the central platform (which was self-initiated except for the first trial in every session); the closing of the doors surrounding the central platform which blocked access to the arms, triggered by entry to the central platform; the opening of the doors after a 5-second delay; and the rat's arrival at the reward location, which probabilistically triggered reward delivery.

Firing rates in both brain areas showed modulation by these timepoints. Firing rates in CA1 showed particular increases prior to exit from the reward platform and around the time of arrival at the reward location (fig. 3.8), with similar firing rate responses to the presence or absence of reward. Firing rates in accumbens also showed increases between exit from the central platform and arrival at the reward location (fig. 3.9).

Cells in the nucleus accumbens have been found to be selective for many elements of a task, including upcoming action choice, predicted action outcome, current action, reward, and reward-prediction error. Although the population variations in firing rates indicate that firing rates were modulated by task events (fig. 3.9), the low trial count in these recordings (average 15 per session) makes it difficult to determine what encoding the accumbens cells might possess with any meaningful statistical power. Anecdotally, many cells showed firing rate patterns that were characteristic but similar across trials regardless of reward outcome or reward location (fig. 3.10A). To compare with previous studies, accumbens cells were divided into "reward-related" and "non-reward-related" by combining all trials in a given session and assessing whether firing rate varied in 250ms bins from the period -1 to +1 second around arrival at the reward location, compared to control time bins (Lansink et al., 2008). Following this method, 23 (13.6%) accumbens cells were reward-related (fig. 3.10B), which is similar to the previously reported figure of 19.8% (Lansink et al., 2008).

Cells in both brain areas have also been reported show theta-modulation, which is believed to be important for routing information from the hippocampus to the accumbens. 60.1% of CA1 and 14.8% of NAc cells showed a theta-modulation index above zero (fig. 3.10C).

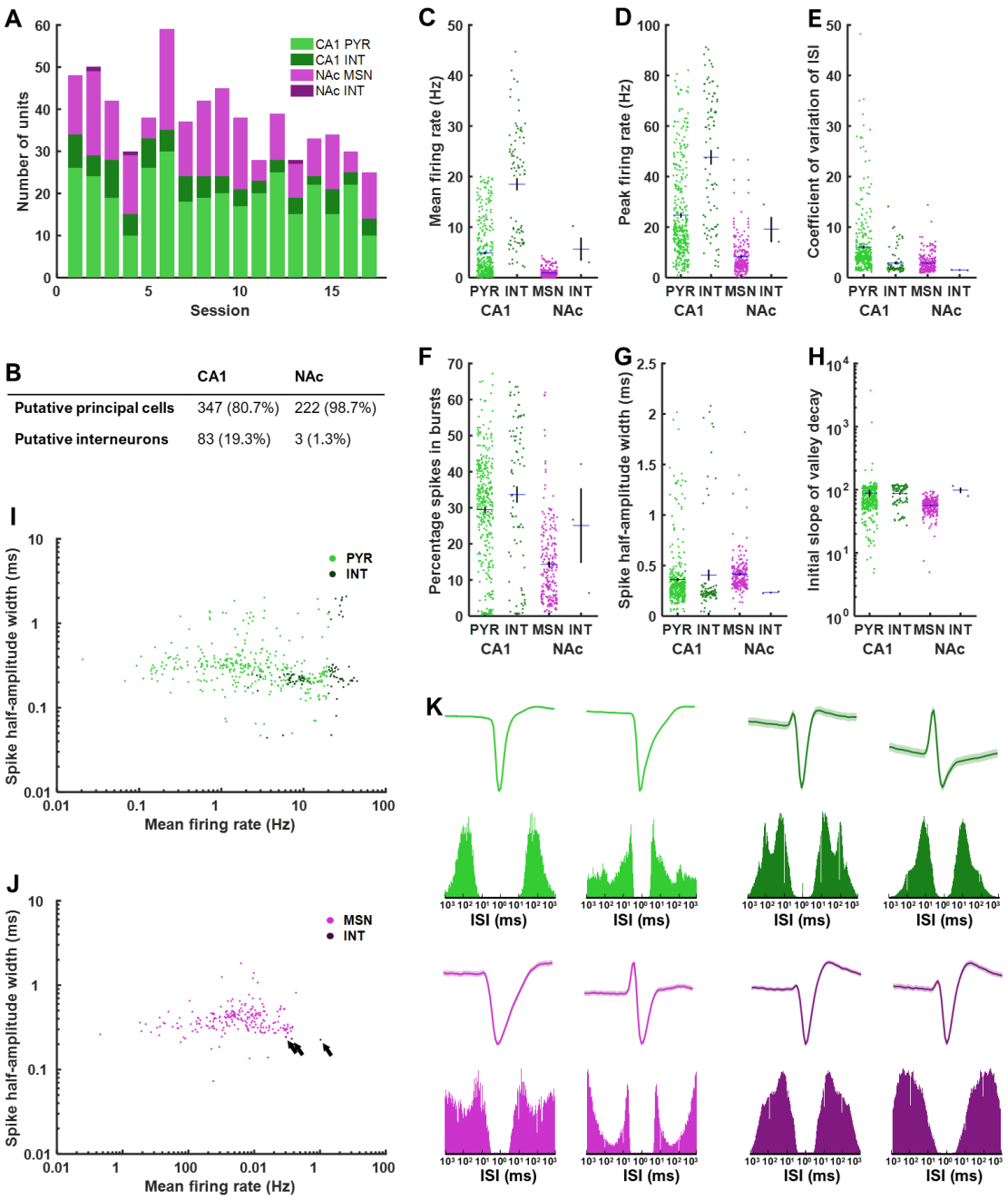

Figure 3.6. Single-unit firing properties. Properties of all 655 units recorded over 17 sessions. **A-B.** Number of units recorded in each session, divided into pyramidal cells (PYR), interneurons (INT) and medium spiny neurons (MSN) and by brain area. **C-H.** Spike waveform and spiking properties of the four cell types over entire recording session, including rest periods. **C.** Mean firing rate. **D.** Peak firing rate: 99th percentile of firing rates convolved in 1-second bins. **E.** Inter-spike interval (ISI) coefficient of variation: ratio of standard deviation to mean. **F.** Burstiness: percentage of ISIs < 15ms. **G.** Spike half-width. **H.** Initial slope of valley decay of spike waveform. **I-J.** Distribution of mean firing rates and spike half-widths: interneurons were defined by high firing rates and mostly had small half-widths. Accumbens interneurons indicated by arrows. **K.** Average spike waveforms and ISI histograms for 8 example neurons.represent binomial confidence intervals. Shaded area in F represents s.e.m. Vertical lines indicate when reward probabilities were reversed.

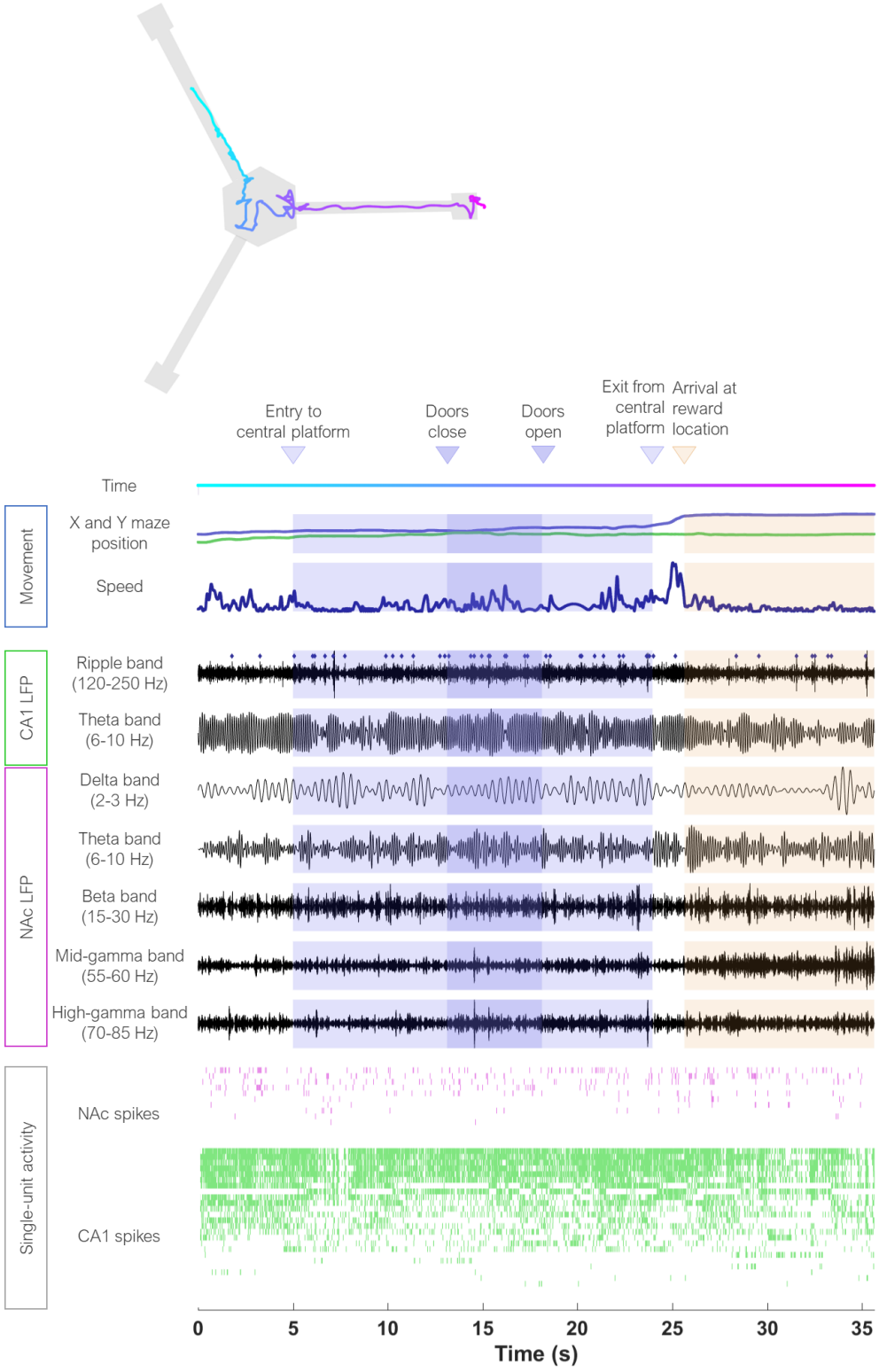

Figure 3.7. Example of behaviour, LFP and spiking activity over one trial. Top: rat's trajectory on maze, starting 5 seconds before entry to the central platform and ending 10 seconds after arrival at the reward location; colour corresponds to "Time" below. Bottom: position on maze, running speed, LFP filtered at specified frequency bands, and spikes. Markers over ripple-band LFP indicate times of discrete ripple events. Shaded areas indicate time on central platform (blue), delay period when doors were closed (dark blue), and time at reward zone (orange).

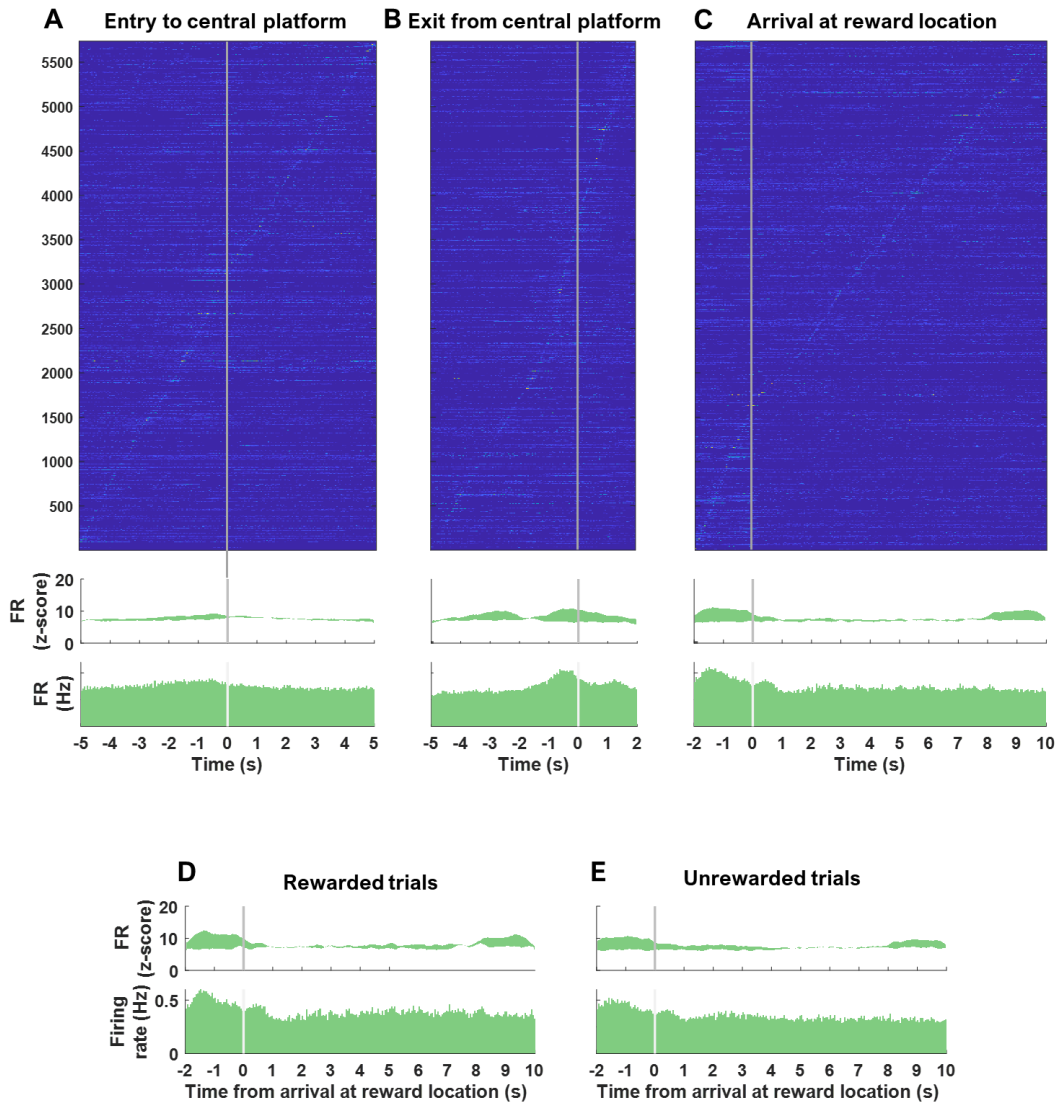

Figure 3.8. Trial-averaged firing rates for CA1. Trial-averaged firing rates of all CA1 cells, aligned to entry (A) and exit (B) from the central platform, and arrival at the reward location (C-E). Top plots show the z-scored firing rate of every cell on every trial in 50ms bins, ordered by the timing of their peak activity. Middle plots show the z-scored firing rate averaged over cells and trials, with standard error. Bottom plots show the raw firing rate averaged over cells and trials.

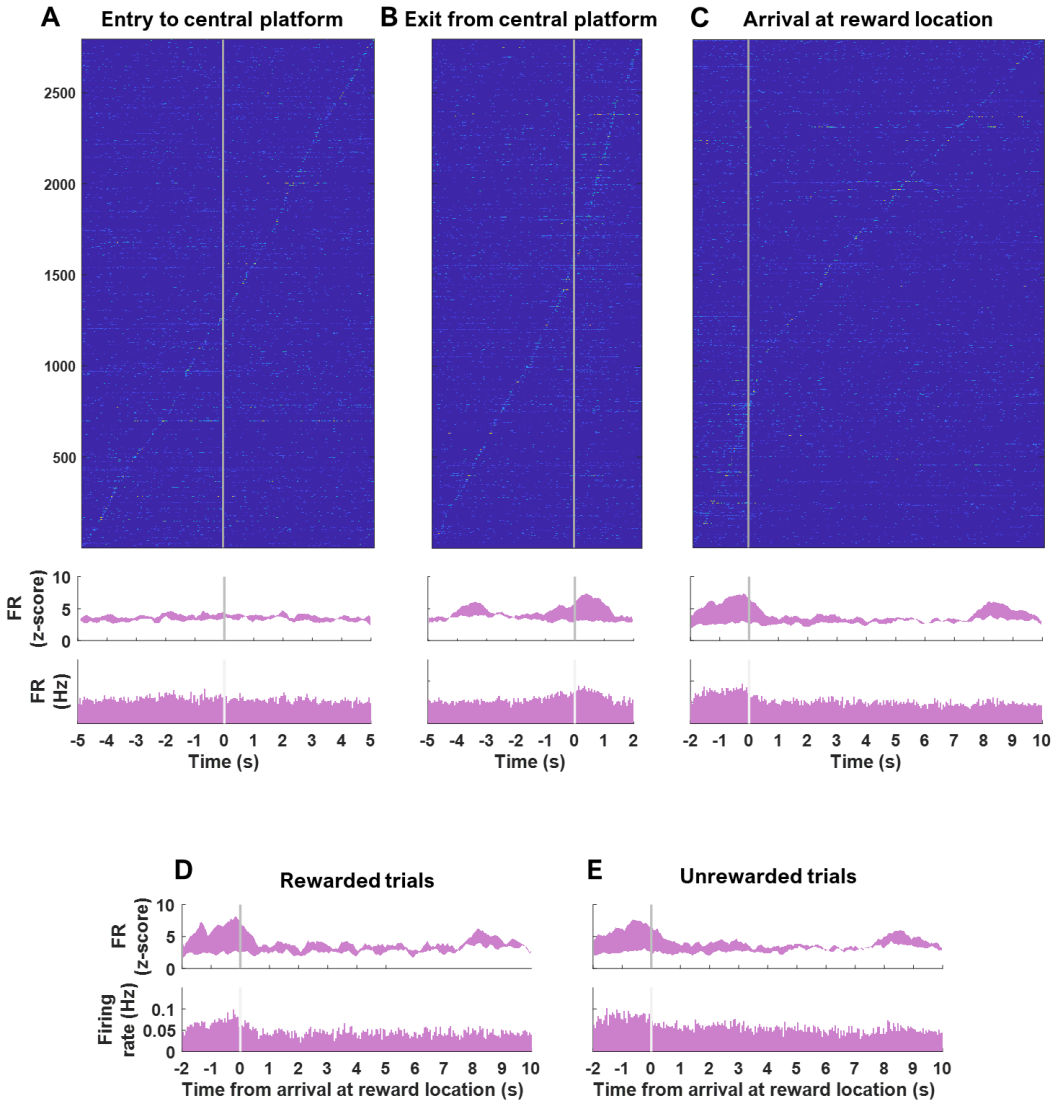

Figure 3.9. Trial-averaged firing rates for accumbens. As figure 3.8: trial-averaged firing rates of all accumbens cells, aligned to entry (**A**) and exit (**B**) from the central platform, and arrival at the reward location (**C-E**). Top plots show the z-scored firing rate of every cell on every trial in 50ms bins, ordered by the timing of their peak activity. Middle plots show the z-scored firing rate averaged over cells and trials, with standard error. Bottom plots show the raw firing rate averaged over cells and trials.

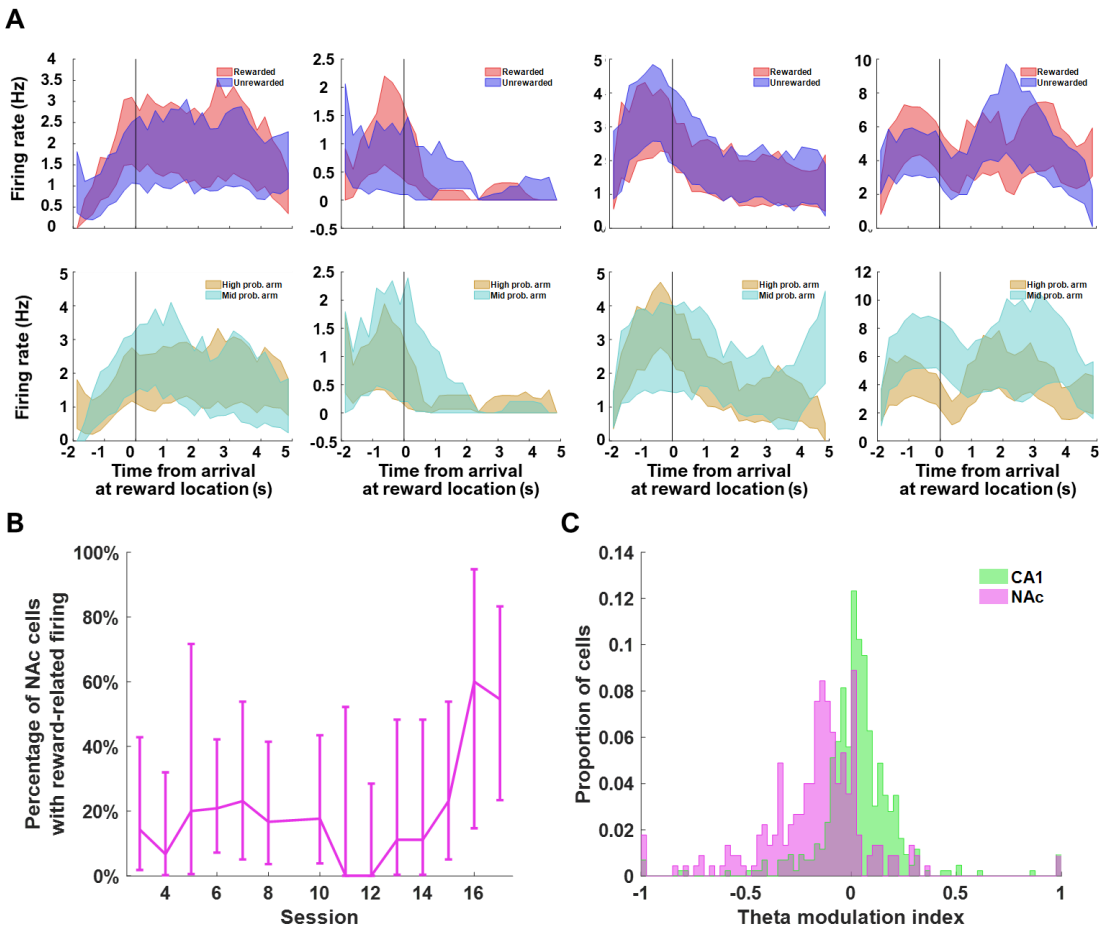

Figure 3.10. Task-related firing of accumbens cells. **A.** Trial-averaged firing rates of four example NAc cells, locked to arrival at the reward location; shaded areas represent mean ± s.e.m. Top row shows averaged firing on rewarded compared to unrewarded trials. Bottom row shows averaged firing on arrivals at the high-probability arm compared to the mid-probability arm. **B.** Percentage of NAc cells for each session which show significantly different firing around arrival at the reward location; error bars indicate binomial confidence intervals. **C.** Histograms of the theta modulation index for CA1 and NAc cells across all sessions.

3.3.3 Behavioural correlates of local field potential

Hippocampal LFP

LFP offers an insight into the information-processing state of the network at large, beyond the sometimes narrow tuning curves and response profiles of individual neurons, serving as a mechanism for communication and synchronisation between large numbers of neurons over long distances. High power in the theta band (6-10 Hz) in hippocampal LFP is strongly associated with running, as well as other active, exploratory behaviours including whisking and vicarious trial-and-error. Accordingly, LFP recorded from

the CA1 probe showed a peak in theta power during approach to the central platform and at arrival at the reward location (i.e. during locomotion), and a dip after arrival at reward location when the rat reliably stopped (fig. 3.11A-B).

Gamma oscillations are frequently observed in CA1, nested in theta cycles during running, and have been linked to memory processes (Colgin & Moser, 2010). Gamma at different frequencies has been suggested to subserve different purposes and route information flow differently through the hippocampus (Zheng et al., 2015; Colgin & Moser, 2010), so mid-gamma (55-60 Hz) and high-gamma (70-85 Hz) oscillations were considered separately. High-gamma power, like theta power, increased during periods of locomotion, showing peaks upon entry to and exit from the central platform (fig. 3.11A), although it was more strongly modulated by the central platform exit and did not exhibit a dip upon reaching the reward location. High-gamma power in CA1 has been associated with higher running speeds (Ahmed & Mehta, 2012; Zheng et al., 2015) and coordination between CA1 and medial entorhinal cortex (MEC; Hafting et al., 2005), and perhaps reflects transmission of spatial information from MEC to CA1 during locomotion, especially the faster runs from central platform to reward location (fig. 3.11B).

Interestingly, hippocampal LFP also showed reward-related correlates. Mid-gamma power exhibited its highest peak 1-4 seconds after arrival at the reward location (fig. 3.11A), likely coinciding either with consumption of the reward or discovery of reward absence, and not associated with locomotion or decision-making. Moreover, mid-gamma power was significantly greater during this period on rewarded trials than unrewarded trials (fig. 3.11E; rank-sum tests, corrected for multiple comparisons). This is not a finding commonly reported in the literature, but may have some link to replay: transient increases in low gamma power have been reported during sharp-wave ripples (Carr et al., 2012), and CA1 cell pairs which co-fire during theta-nested mid-gamma oscillations are preferentially reactivated during ripples (Lopes-dos-Santos et al., 2018), although the mid-gamma increase here coincided with low theta power. This dip in theta power was itself modulated by reward, being larger on rewarded trials (fig. 3.11C), which may reflect different motor patterns, such as reward consumption, whisking and rearing, upon discovering the presence or absence of reward (Wyble et al., 2004). Indeed, this coincided with less movement on rewarded trials (fig. 3.11G). Differences in hippocampal LFP between rewarded and unrewarded trials are unlikely to stem from noise artefacts relating to the consumption of reward, because rewards were liquid (prompting licking, which causes less vibration than chewing) and there were no obvious licking-related noise artefacts dominating the raw LFP (fig. 3.11I-J).

Concurrent with theta power peaks were peaks in beta (15-30 Hz; fig. 3.11A) power, a harmonic of theta. Beta power was particularly high during the reward location approach, consistent with previous reports of increased beta power during reward expectancy (Lansink et al., 2016), but there was no difference in beta power between trials where reward expectancy can be assumed to be high compared to trials where it can be assumed to be lower (fig. 3.11H; rank-sum tests). Specifically, from session 5 the rat's performance

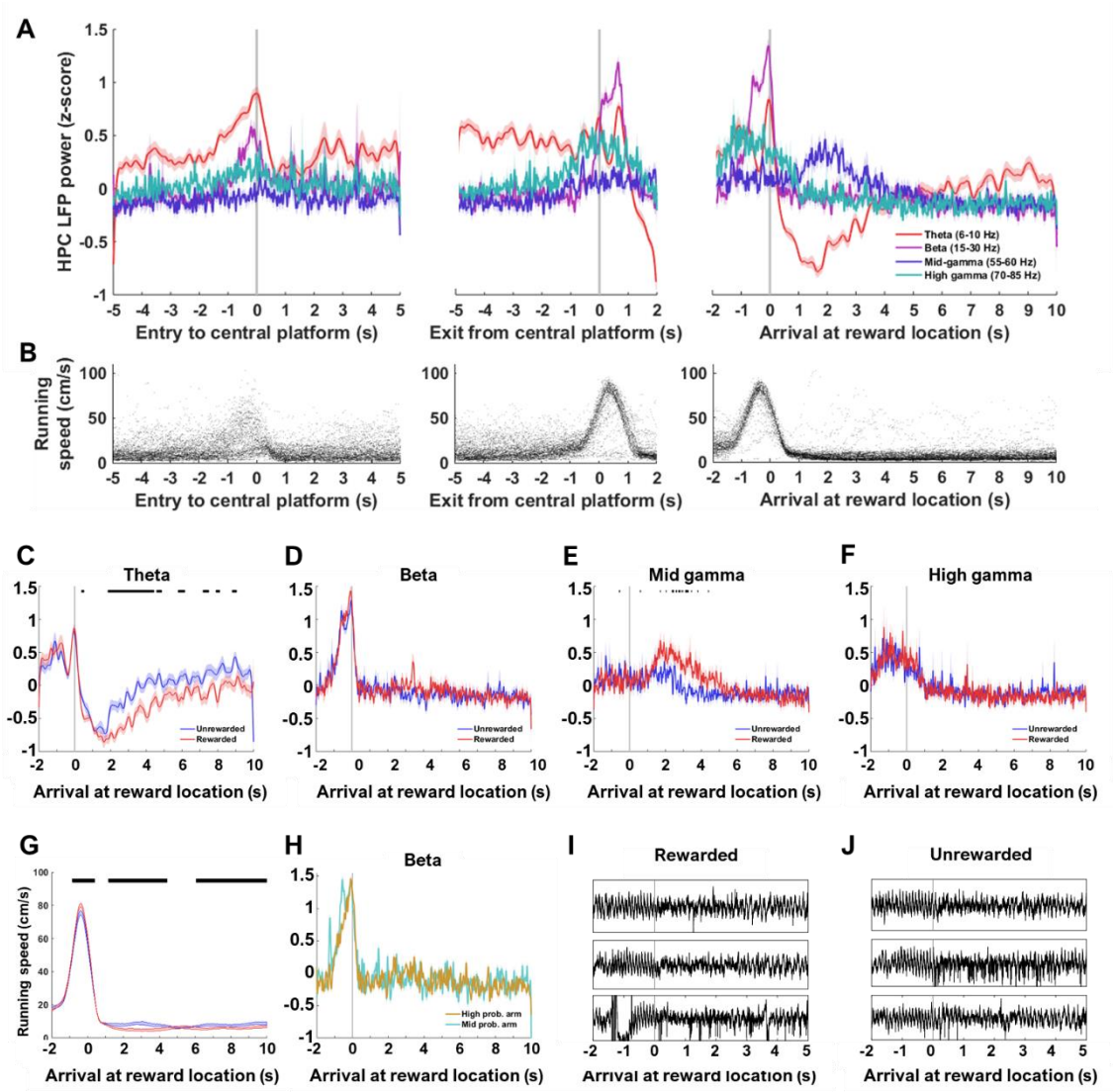

Figure 3.11. Trial-averaged LFP power recorded from CA1 probe. A. Power at theta, beta, mid gamma and high gamma frequencies, averaged over all trials. B. Running speed, smoothed with a window of 1s, for every trial. C-F. Trial-averaged power for rewarded and unrewarded trials; time bins where power is significantly different is indicated in black, calculated using rank-sum tests and corrected for multiple comparisons using the Benjamini-Hochberg method with a false discovery rate of 0.05. G. Trial-averaged running speed, smoothed with a window of 1s, for rewarded and unrewarded trials; black indicates statistical significance as above. H. Trial-averaged beta power for trials on which rat entered the high-probability arm and mid-probability arm, irrespective of reward outcome. I-J. Example raw hippocampal LFP trace from three example rewarded (I) and three unrewarded (J) trials; no systematic noise differences were apparent.

was above chance (fig. 3.4E) so it can be assumed that its knowledge of reward probabilities was reasonably accurate from sessions 5 to 12, after which reward probabilities were changed. Beta power during approach to the high-probability arm in these sessions was no different from that during approach to the mid-probability arm (fig. 3.11H), suggesting that it is not scaled to the degree of reward expectancy.

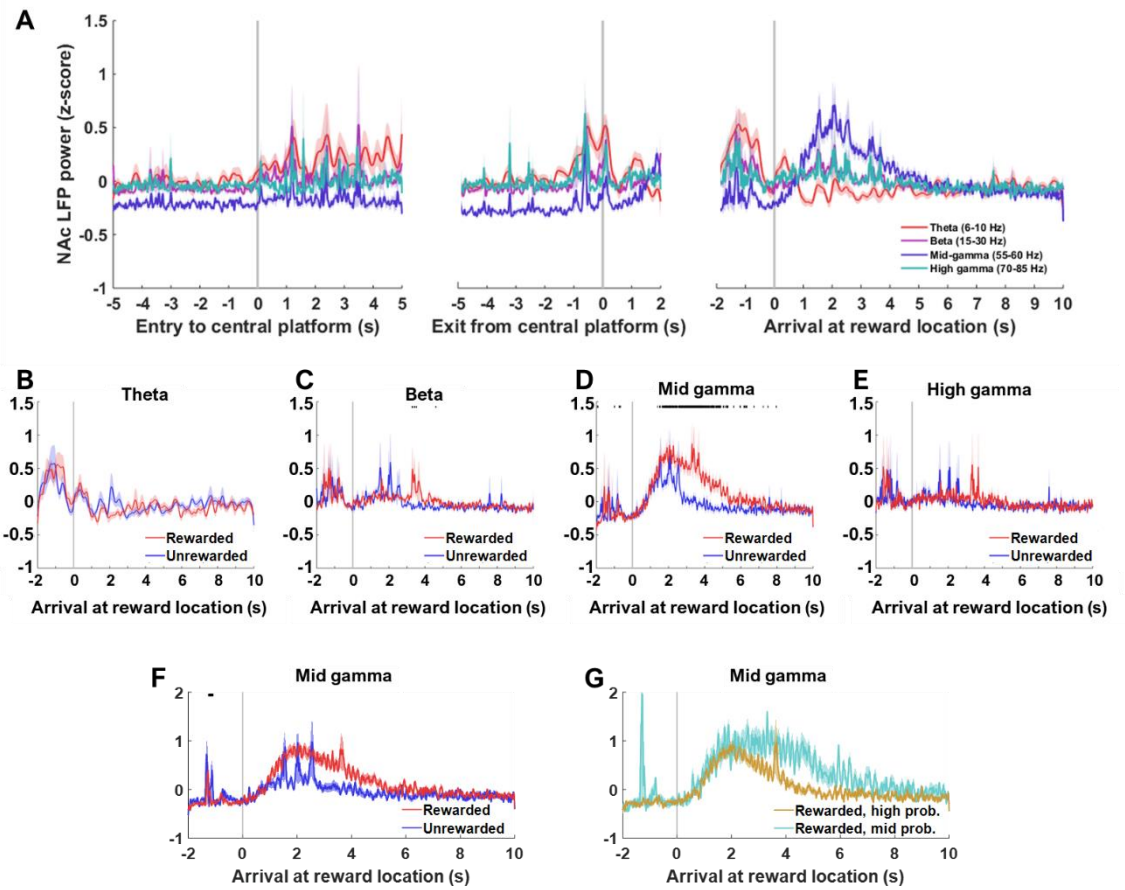

Figure 3.12. Trial-averaged LFP power recorded from accumbens probe. **A.** Power at theta, beta, mid gamma and high gamma frequencies, averaged over all trials. **B-E.** Trial-averaged power for rewarded and unrewarded trials; time bins where power is significantly different is indicated in black, calculated using rank-sum tests and corrected for multiple comparisons using the Benjamini-Hochberg method with a false discovery rate of 0.05. **F.** As D, but calculated over trials in sessions 5-12. **G.** As D, but calculated over rewarded trials at the high- and mid-probability arms in sessions 5-12.

Accumbens LFP

Previous studies have found that gamma power in the accumbens is modulated at various stages of action initiation, execution and outcome. Increases in mid-gamma power have been found at the point of movement initiation (van der Meer & Redish, 2009), and during anticipation of a reward-instructive cue (Donnelly et al., 2014), while high-gamma power has been reported to increase as an animal approaches a reward location (van der Meer & Redish, 2009). Following reward, there are conflicting reports of increases (van der Meer & Redish, 2009) and decreases (Berke, 2009) in mid-gamma power, as well as inconsistent reports of increases in high-gamma power (Berke, 2009; van der Meer et al., 2019). These discrepancies might arise because the gamma response to task variables changes over the course of learning (van der

Meer & Redish, 2009), or because the gamma response is very localised and depends on precisely where recordings are made, or because these oscillations arise non-locally and might reflect processing of other kinds of task-dependent information in cortex (Carmichael et al., 2017; Berke, 2009). Alternatively, accumbens gamma oscillations might be something of a neural red herring, reflecting a default network state when the animal is at rest rather than information processing (Malhotra et al., 2015).

In these results, mid-gamma power showed a small increase prior to exiting the central platform, perhaps reflecting movement initiation, and a larger increase after arrival at the reward location (fig. 3.12A). This was reward-dependent, as it was higher on rewarded than unrewarded trials (fig. 3.12D, rank-sum tests). As accumbens has been found to respond to both reward and reward-prediction error, this reward-responsive increase in mid gamma was probed further. Following the logic that the rat had good approximate knowledge of reward probabilities at each arm in sessions 5 to 12, we can assume that RPE was greater following reward at the mid-probability arm than the high-probability arm in those sessions. Mid-gamma power appeared to be greater following reward than non-reward in sessions 5-12, consistent with other sessions, although with the smaller sample size of trials this difference fell below statistical significance (fig. 3.12F). Dividing the rewarded trials into those at the high-probability and mid-probability arms (and discarding those at the low-probability arm) revealed that both levels of reward-prediction error elicited a similar magnitude of peak in mid-gamma power, but a non-significant tendency for the mid gamma increase to persist for 2-3 seconds longer following a greater reward-prediction error (fig. 3.12G). The non-significance of this result cautions against over-interpretation, but it is more consistent with the proposal that gamma oscillations reflect reward processing than the suggestion of a default rest state or association with olfaction (Malhotra et al., 2015; Carmichael et al., 2017): the behavioural state (e.g. licking, resting) and olfactory input can be assumed not to differ between rewarded trials on the high- and mid-probability arms.

High-gamma power, meanwhile, showed some peaks around central platform exit and after arrival at the reward location, but little consistency in its power and scant evidence for the reported ramping increase towards reward (fig. 3.12A). Theta power was highest following entry to the reward platform – during periods of exploration and vicarious trial-and-error seen in the cohort of rats in Chapter 2 – and upon exit from the reward platform, reflecting movement initiation, before decreasing as the rat approaches the reward location (fig. 3.12A-B). Beta power showed some variability during reward-location approach but little consistent modulation (fig. 3.12A & C).

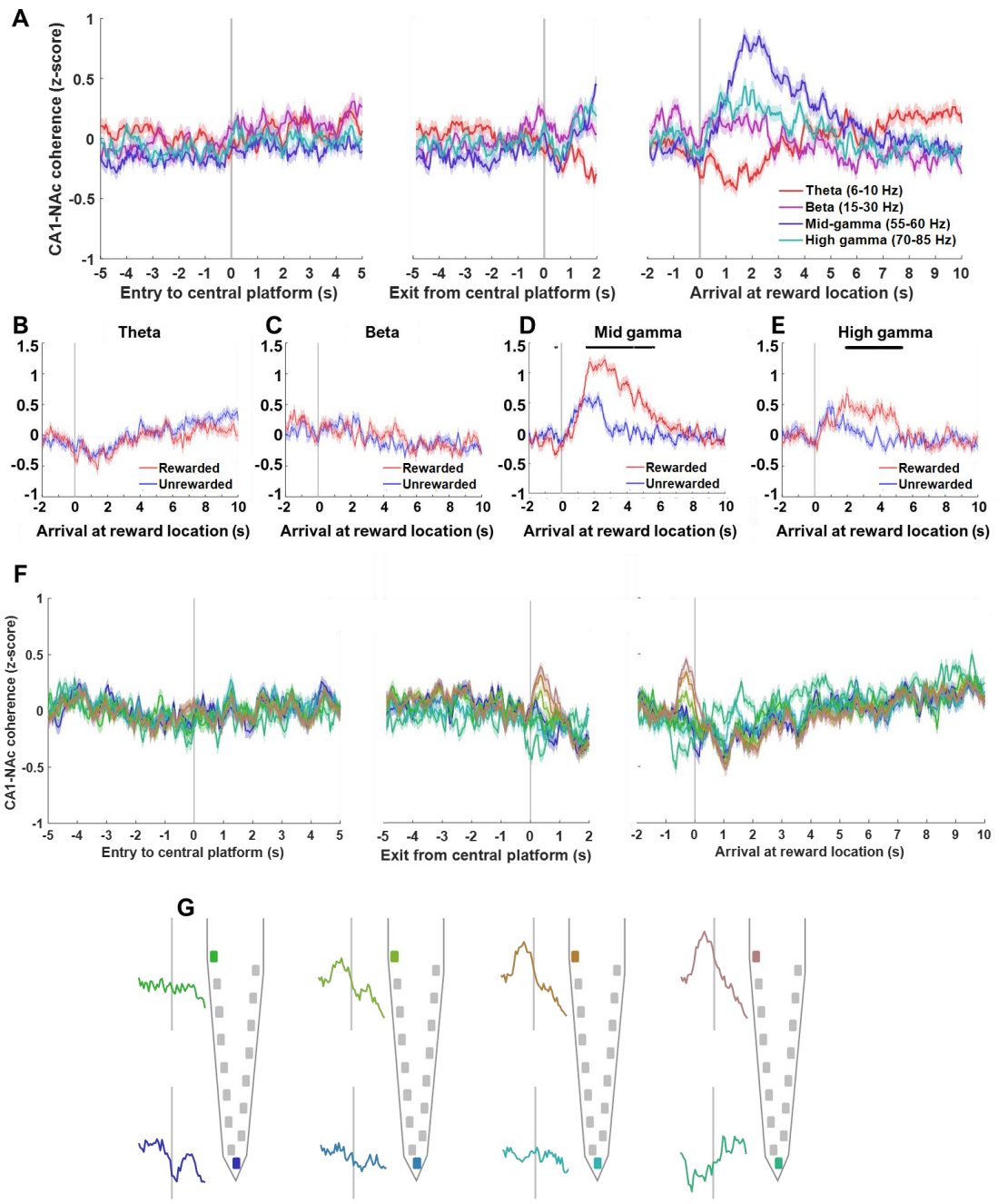

Figure 3.13. Trial-averaged coherence between CA1 LFP and NAc LFP. A. Coherence at theta, beta, mid gamma and high gamma frequencies, averaged over all trials. B-E. Trial-averaged coherence for rewarded and unrewarded trials; time bins where power is significantly different is indicated in black, calculated using rank-sum tests and corrected for multiple comparisons using the Benjamini-Hochberg method with a false discovery rate of 0.05. F. Trial-averaged coherence in the theta band between CA1 and 8 NAc recording sites; colours correspond to the sites illustrated in G. G. Recording sites on the NAc probe from where coherence was analysed, with theta coherence plotted from -1 to 1 second from arrival at reward location.

3.3.4 Hippocampus-accumbens LFP coherence

The results presented so far show that there were responses in the LFP of both areas to task variables including locomotion (CA1 theta, CA1 high gamma), movement initiation (NAc theta), reward approach (CA1 beta), and reward outcome (CA1 mid gamma & NAc mid gamma). Results from other studies have found LFP synchronisation in the theta and beta bands associated with goal-directed navigation (Lansink et al., 2016), through which the hippocampus is proposed to influence motor output.

Here, coherence in neither theta nor beta bands showed a substantial increase throughout the trials (fig. 3.13A), in keeping with the previous finding that uncued goal-directed navigation does not elicit such coherence (Lansink et al., 2016). However, there was a decrease in theta coherence during the reward approach and after arrival at the reward location, in direct opposition to the previous finding of a "ramping" increase in theta coherence leading up to arrival at a reward location (van der Meer & Redish, 2011; fig. 3.13A). This could be due to anatomical variations in the density of hippocampal axonal targets in the accumbens (Voorn et al., 2004; Pennartz et al., 2011; Trouche et al., 2019), so LFP coherence was obtained between the same CA1 channel and 8 accumbens channels to compare, from the top and bottom of each shank (fig. 3.13G). Theta coherence during reward approach was dependent on recording location, with ramping up apparent in the more dorsolateral area of recording (fig. 3.13F-G). This was despite no such topological variation in trial-averaged theta power at the same accumbens channels (data not shown).

Given the reward-related changes in gamma power in both brain areas, coherence in the mid-gamma and high-gamma ranges were also examined. To my knowledge, hippocampal-accumbens coherence in the gamma range has not been reported in the literature, but there was a post-reward-outcome increase in this task in both mid-gamma coherence and high-gamma coherence (fig. 3.13A), and moreover this coherence was greater in rewarded trials than unrewarded. (fig. 3.13D-E).

3.3.5 Sharp-wave ripples

Sharp-wave ripples, transient bursts of high-frequency oscillations in the 120-250 Hz range, are a feature of hippocampal LFP commonly seen during periods of slow-wave sleep or wakeful rest, and strongly associated with replay events. Having established so far that features of single-unit and local field activity in both CA1 and accumbens encode task-related information, to relate this to replay I investigated whether and when ripples occurred on the maze and how they modulated activity in both areas.

During wake, ripples are thought to occur primarily after the hippocampus switches from an online state, dominated by incoming sensory information and encoding current spatial information, to an offline state

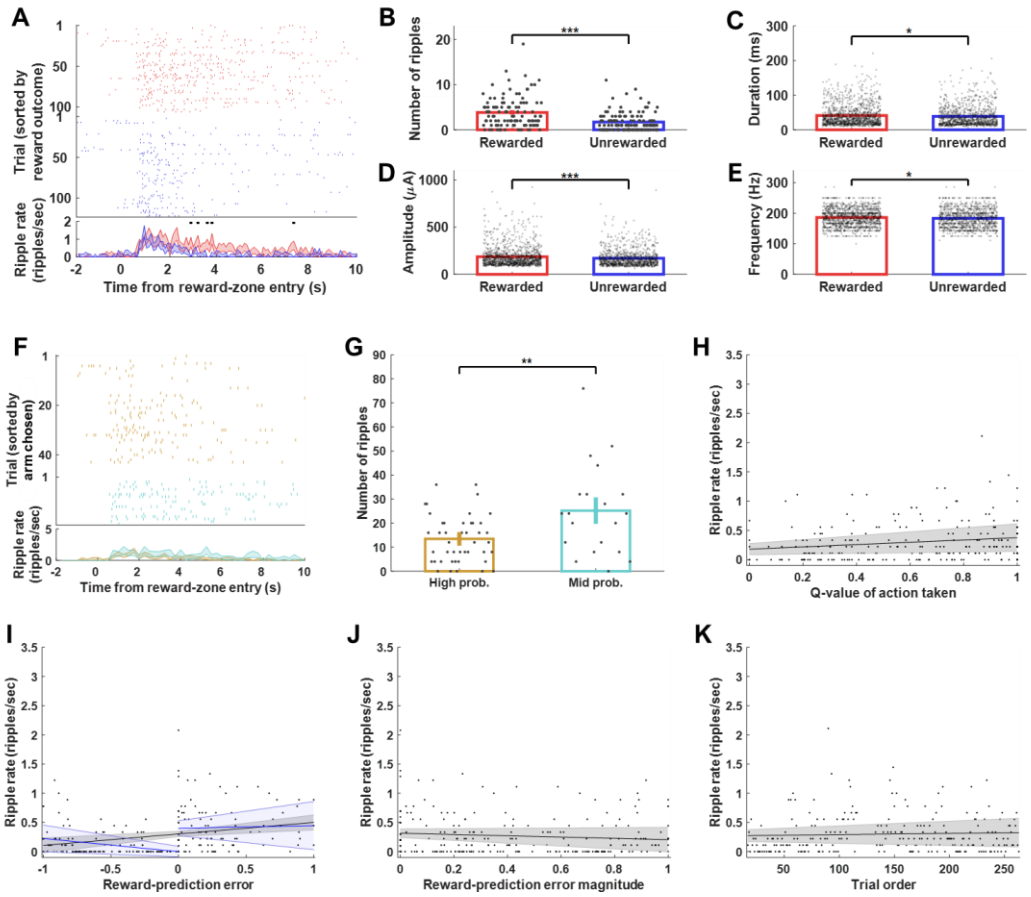

Figure 3.14. Sharp-wave ripples. **A.** Top: ripple events from every trial, aligned to arrival at reward location, divided into rewarded and unrewarded trials. Bottom: trial-averaged ripple rate in bins of 100ms for rewarded and unrewarded trials; black indicates significant differences, calculated using rank sum tests and corrected for multiple comparisons using the Benjamini-Hochberg method with a false discovery rate of 0.05. **B-E.** Comparison of ripple properties in the 1-10 seconds after arrival at the reward location. * indicates significant difference at $p < 0.05$; *** indicates significant differences at $p < 0.001$; calculated using rank sum tests and Bonferroni-corrected. **F.** Ripples occurring on rewarded trials at the high- and mid-probability arms, on sessions 5-12. **G.** Number of ripples at 1-10 seconds on rewarded trials at the high- and mid-probability arms; ** indicates significant difference at $p < 0.01$, calculated using one-tailed paired t-test. **H.** Regression of ripple rate at 1-10 seconds against estimated Q-value of action chosen on each trial. **I.** Ripple rate 1-10 seconds after arrival plotted against the estimated reward-prediction error (RPE) for each trial. Black line shows regression for all trials; blue lines show regression for trials with positive and negative RPE respectively; shaded areas represent confidence intervals. **J.** Regression of ripple rate at 1-10 seconds against RPE magnitude, i.e. absolute RPE. **K.** Regression of ripple rate at 1-10 seconds against trial order, i.e. experience of the task.

dominated by spontaneously generated activity when the animal is at rest, including during periods of grooming or feeding (Buzsáki, 2015). Higher ripple rates have previously been found following reward than no-reward on a maze task (Singer & Frank, 2009; Ambrose et al., 2016; Sosa et al., 2019), purportedly

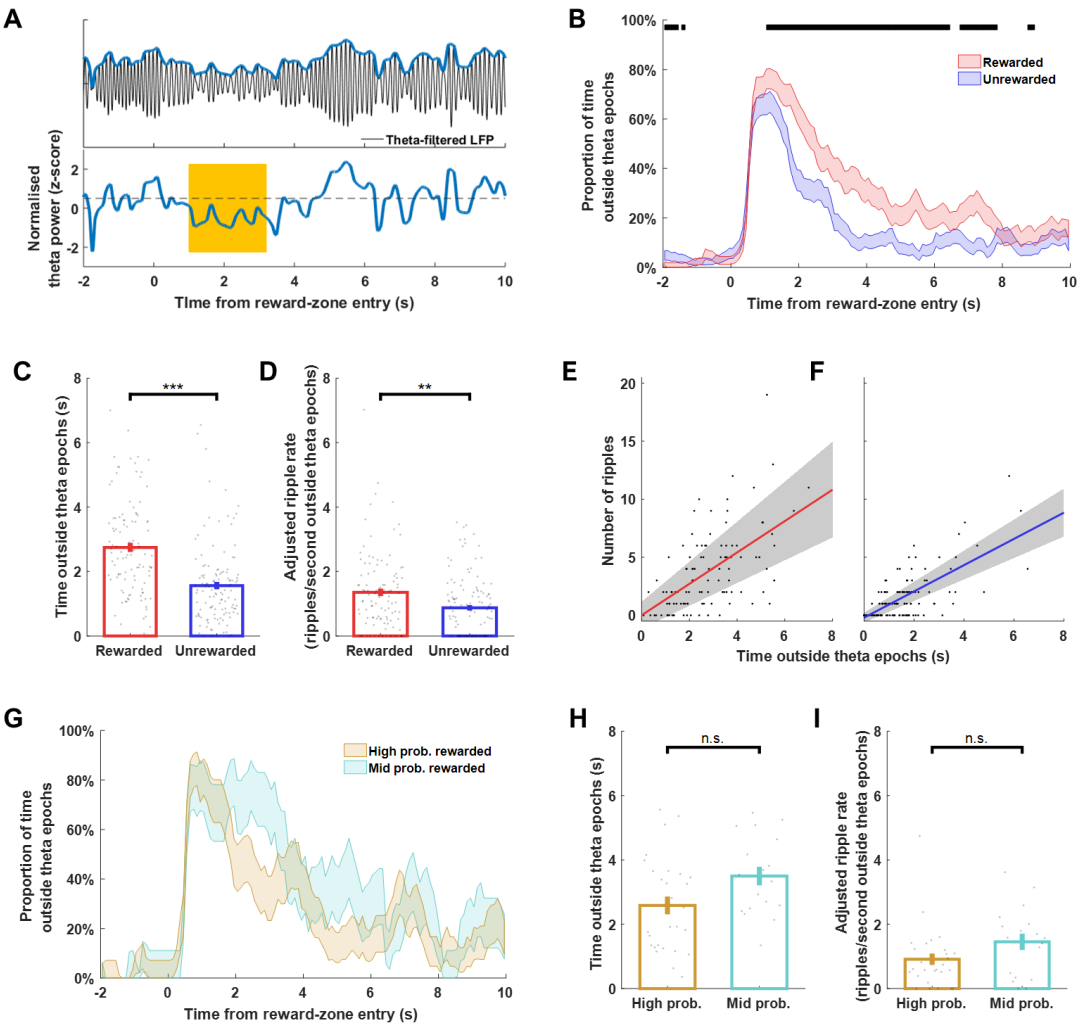

Figure 3.15. Adjusted sharp-wave ripple rates. **A.** Top: an example trial (unrewarded), showing hippocampal LFP filtered at theta frequency (6-10 Hz) around the time of arrival at reward location. Blue line shows theta power envelope, which was reduced roughly 1-4 seconds after arrival. Bottom: Theta power envelope from the same trial, normalised with reference to theta power over the whole TASK period. Ripple detection was restricted to periods where normalised theta power fell below 0.5 for at least 1 second; this non-theta epoch is indicated in yellow, lasting 2.2 seconds. **B.** The proportion of time which fell outside of theta epochs (i.e. the proportion of time which was "eligible" for ripple detection), averaged over rewarded and unrewarded trials. Black bars indicate significant differences between rewarded and unrewarded trials ($p<0.05$, Benjamini-Hochberg corrected, false discovery rate 0.05). **C.** Average time outside theta epochs in the 1-10 seconds after arrival at reward location. **D.** Number of ripples per second outside theta epochs in the 1-10 seconds after arrival at reward location. **E-F.** Correlation between time outside theta epochs on rewarded (E) and unrewarded(F) trials and number of ripples 1-10 seconds after arrival at reward location. **G.** As B, for rewarded trials at the high- and mid-probability arms. **H-I.** As E-F for rewarded trials at the high- and mid-probability arms.

promoting greater replay of rewarded than unrewarded information. In agreement with these previous reports, the rate of ripples was very low prior to arrival at the reward location (during running), and increased sharply after arrival (fig. 3.14A). The rate of ripples was higher following a rewarded outcome

than an unrewarded outcome (fig. 3.14A-B), which might reflect a longer period of immobility while reward is consumed on rewarded trials, or increased excitability in the hippocampus in response to reward (or both). In addition to ripples following reward being more numerous, they were also slightly longer in duration (fig. 3.14C) and reached a slightly higher peak frequency (fig. 3.14E), although these differences were very small, and reached a higher peak amplitude (fig. 3.14D). Dopamine administration to CA1 has been found to increase the rate and amplitude of ripples in CA1 (Miyawaki et al., 2014), so these effects could reflect increased dopamine release in CA1 in response to reward compared to no-reward.

This CA1 dopamine release would likely originate from the ventral tegmental area (VTA), whose dopaminergic axon terminals fire phasically in CA1 (Lisman & Grace, 2005), signalling a positive reward-prediction error (Schultz & Dickinson, 2000). If this is the case, more ripples (reflecting dopamine-triggered increases in CA1 excitation) should be seen in response to rewards that are more unexpected. Following the logic that the rat had good approximate knowledge of reward probabilities at each arm in sessions 5 to 12, we can assume that RPE was greater following reward at the mid-probability arm than the high-probability arm on those sessions. Accordingly, there were more ripples following reward at the mid-probability arm than the high-probability arm (fig. 3.14F-G; one-tailed paired t-test, $p = 0.0077$). This relationship was not robust, however, when the association was assessed more precisely: RPE was estimated according to Q-value estimates using methods outlined in Chapter 5. Although ripple rate was significantly correlated with RPE overall (fig. 3.14H, Pearson's correlation coefficient = 0.30, $p = 4.6 \times 10^{-6}$), it was not correlated with either positive RPE (fig. 3.14H, $p = 0.76$), or overall RPE magnitude (fig. 3.14I, $p = 0.11$). This suggests that after controlling for reward outcome (reward or no-reward), there is no relationship between RPE and ripple rate. However, there was a significant correlation between negative RPE and ripple rate (Pearson's correlation coefficient = -0.29, $p = 0.0015$), which suggests that particularly surprising absence of reward, as well as presence of reward, increases ripple rate. The overall picture of how ripple rates are influenced by reward and reward-prediction error, therefore, is inconsistent.

To resolve some of this ambiguity, the effect of immobility on detected ripple rate was controlled for. Sharp-wave ripples are produced in the hippocampus when theta power is low (Buzsáki et al., 1983), so by default ripples during TASK were included only if they occurred outside theta epochs (see 3.2.5 Methods for details of ripple detection). Theta power therefore not only serves as a proxy for behavioural state, but also restricts by definition how many ripples could pass the criteria for inclusion: the example unrewarded trial in fig. 3.15A shows a short decrease in theta power upon reaching the reward location before it increases again; correspondingly, there was a period (shown in orange) of just 2.2 seconds in which theta power fell low enough for ripples to take place by definition. This means just 24.4% of the total time from 1 to 10 seconds after arrival at the reward location in the example trial was "eligible" for ripples. Across all trials, there was significantly more eligible low-theta time following reward than no-reward (fig. 3.15B-C; rank-sum test, $p = 1.3 \times 10^{-10}$, Bonferroni-corrected), and the number of ripples was consequently correlated with the amount of time outside theta epochs (fig. 3.15E-F. Rewarded: Pearson's correlation coefficient =

0.58, $p < 0.0001$. Unrewarded: Pearson's correlation coefficient = 0.70, $p < 0.0001$). To control for this, the ripple rate in fig. 3.14B was adjusted to reflect ripples per eligible low-theta time, not total ripple numbers. The ripple rate was still higher on rewarded than unrewarded trials (fig. 3.15D; rank-sum test, $p = 0.0022$, Bonferroni-corrected), consistent with the results for non-adjusted ripple rate (fig. 3.14B). The same adjustment was made for ripple rate following reward on high- and mid-probability arms; no significant differences were found between them in eligible low-theta time (fig. 3.15G-H) or adjusted ripple rate (fig. 3.15I). Taken together, these results are consistent with prior findings in the literature that ripple rate scales with reward and not reward-prediction error (Singer & Frank, 2009; Ambrose et al., 2016; Sosa et al., 2019).

3.3.6 Modulation of single-unit firing rates by sharp-wave ripples

During a sharp-wave ripple, a substantial minority of hippocampal cells increase their firing rate. The recruitment of cells during ripples often reflects ensembles that are co-active during behaviour: most notably, sequences of place cell spikes that encode recent trajectories occur again, compressed in time, during ripples. Over many ripples, the effect is that some cells show modulation of their firing rates around the time of ripple events, and because of the link to replay, cells which are ripple-modulated may encode the replayed information during behaviour. Moreover, an association has been found between cells which encode an aspect of a trial (e.g. place cells active during a run) and reactivation during ripples immediately after the event, for example after receipt of a reward (Singer & Frank, 2009), so the activity of cells during ripples which occur during or at the end of a trial may hold clues to what is being replayed. This is particularly useful if reward, or high reward-prediction error, induces dopamine release in CA1, as it may modulate the plasticity at active synapses during this period.

During ripples, replay events are observed not only in the hippocampus but in a range of other cortical and subcortical areas, including the nucleus accumbens. A substantial fraction of accumbens cells are also reported to be modulated by hippocampal ripples (Pennartz et al., 2004; Sosa et al., 2019). This is purported to be the mechanism by which hippocampal replay is "broadcast" to the accumbens for relevant information-encoding cells to be recruited into the replay event; consistent with this idea is the finding that hippocampal cells encoding place information and accumbens cells encoding reward information are reactivated concurrently during ripples (Lansink et al., 2009; Sjulson et al., 2018).

The results here corroborate the finding that a substantial fraction of cells in both areas are modulated by hippocampal ripples, during the TASK period as well as PRE- and POST-task rest periods (fig. 316A-B). Specifically 90.0% of CA1 cells and 84.0% of accumbens cells were significant activated and/or suppressed during ripples during at least one of these period (Table 3.1).

	CA1			Accumbens		
	Positively modulated	Negatively modulated	Unmodulated	Positively modulated	Negatively modulated	Unmodulated
PRE	69.8%	13.7%	16.5%	45.8%	17.8%	36.4%
TASK	56.5%	12.6%	30.9%	32.0%	18.7%	49.3%
POST	69.5%	14.0%	16.5%	38.2%	25.3%	36.4%
Any	79.8%	24.2%	36.7%	66.2%	39.6%	68.4%

Table 3.1. Ripple-modulation of firing rates. Proportion of cells in CA1 and NAc which were significantly modulated by ripples during the TASK period and PRE- and POST-task rest, for sessions 1-17.

CA1 cells whose firing rates were positively modulated by ripples during TASK had a higher firing rate than other CA1 cells (fig. 3.16C; rank-sum test, $p = 1.4 \times 10^{-14}$) and a marginally lower theta modulation index (fig. 3.16F; paired t-test, $t = 2.8$, $p = 0.019$; all Bonferroni-corrected). This difference in theta modulation suggest that accumbens cells which were ripple-modulated during TASK might engage more during the run, or

show greater connectivity with CA1 during the run, when theta coherence between hippocampus and accumbens was at its highest (fig. 3.11F).

To examine when these cells were active on the maze, trial-averaged firing rates were obtained and compared for subsets of cells defined by their ripple-modulation during TASK. CA1 cells which were positively ripple-modulated showed an increase in firing rate around the time of exit from the central platform (fig. 3.16G), peaking on the central platform around the time of movement initiation. CA1 cells which were not ripple-modulated showed a steadier firing rate throughout. The biggest modulation in firing rate was observed in negatively modulated cells, which showed a steady decrease in firing rate until arrival at the reward location, at which point it rebounded. The association between modulation by ripples and firing throughout the task suggests that cells which carry spatial information about the trajectory (which increase their firing rate between the times of central platform exit and reward location arrival) are preferentially recruited to ripples during TASK, while those which do not carry spatial information about the trajectory (which show a persistent decrease in firing throughout locomotion) are suppressed during ripples. This, in turn, is indicative of replay of task-relevant activity in the hippocampus.

In contrast to CA1, ripple-modulation of accumbens cells had a much looser association with firing during the trial. All subsets of accumbens cells showed increases in firing rate between the point of exit from the central platform and arrival at the reward location (fig. 3.16H). There was some tendency for positively modulated cells to peak earlier, immediately after central platform exit, and unmodulated cells to peak later,

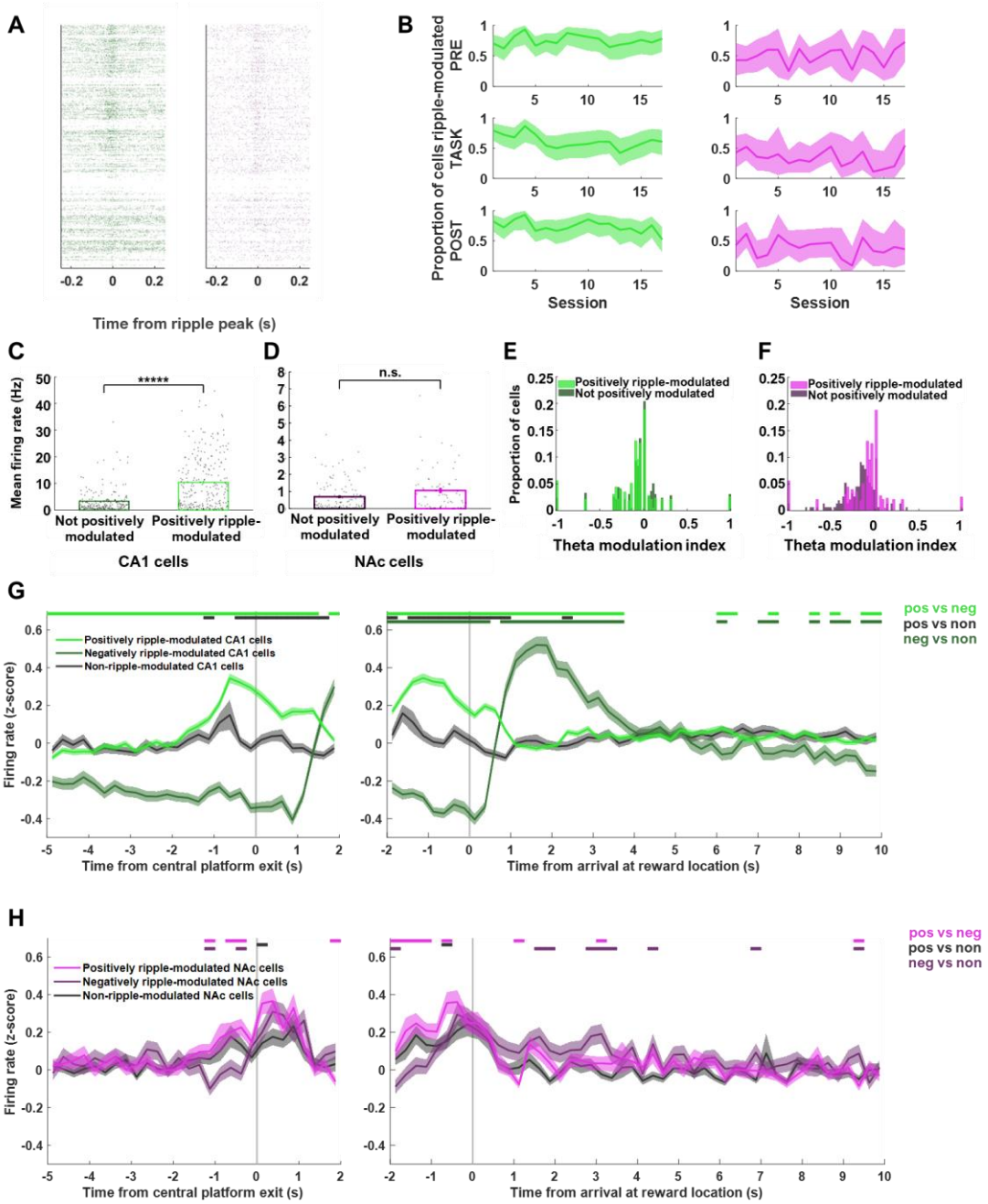

Figure 3.16. Firing properties of ripple-modulated cells. **A.** Ripple-triggered spiking of a ripple-modulated CA1 cell (left) and a ripple-modulated NAc cell (right). **B.** Proportion of CA1 cells (left) and NAc cells (right) which were positively modulated by ripples during PRE (top), TASK (middle row) and POST (bottom). **C.** Mean firing rate of positively modulated CA1 cells and negatively or unmodulated CA1 cells. **D.** Mean firing rate of positively modulated NAc cells and negatively or unmodulated NAc cells. **E.** Distribution of theta modulation indices (TMI) for positively ripple-modulated and unmodulated CA1 cells. **F.** Distribution of theta modulation indices (TMI) for positively ripple-modulated and unmodulated NAc cells. **G.** Trial-averaged z-scored firing rates of positively ripple-modulated and unmodulated CA1 cells; black indicates significant differences. **H.** Trial-averaged z-scored firing rates of positively ripple-modulated and unmodulated NAc cells; no significant differences.

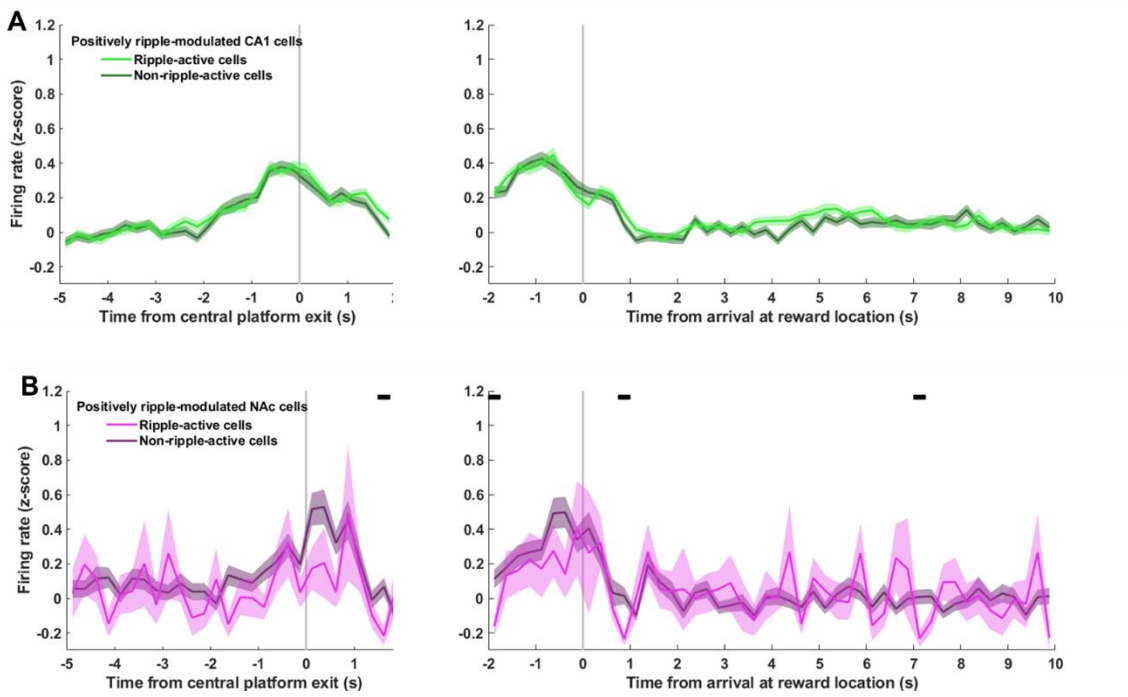

Figure 3.17. Firing rates of positively ripple-modulated cells. **A.** Trial-averaged z-scored firing rates of positively ripple-modulated CA1 cells, divided for each trial into cells that fired at least one spike during a ripple following arrival at reward location, and cells that did not; no significant differences. **B.** Trial-averaged z-scored firing rates of positively ripple-modulated NAc cells, divided for each trial into cells that fired at least one spike during a ripple following arrival at reward location, and cells that did not; black indicates significant differences.

upon reward location arrival, but this trend was dwarfed by the much stronger pattern of overall firing rate increases during locomotion. This suggests that cells which encode spatial information along the trajectory from central platform to reward location show only a slight preference in firing around the time of ripples during TASK.

Ripples during a task have been shown to impact spatial learning (Jadhav et al., 2012), and to bias their content towards trajectories which include an animal's current location (Ólafsdóttir et al., 2017), especially following rewarded trials (Singer & Frank, 2009). So to further examine what activity might be encoded during ripples immediately following reward outcome, positively modulated cells were divided on each trial into those which fired a spike during at least one ripple and those which did not. A higher firing rate during the trial amongst cells which subsequently took part in ripples would indicate replay of the trajectory just taken; cells which were not positively modulated by ripples (and which therefore would be less likely to take part in ripples) were excluded, and spikes which occurred during ripples were also excluded as these would, by definition, only occur for the ripple-active group. In CA1 there was no difference: both subgroups showed an increase in firing which peaked prior to central platform exit, and subsequent participation in ripples was not associated with any firing rate differences (fig. 3.17A). In accumbens, the pattern was very

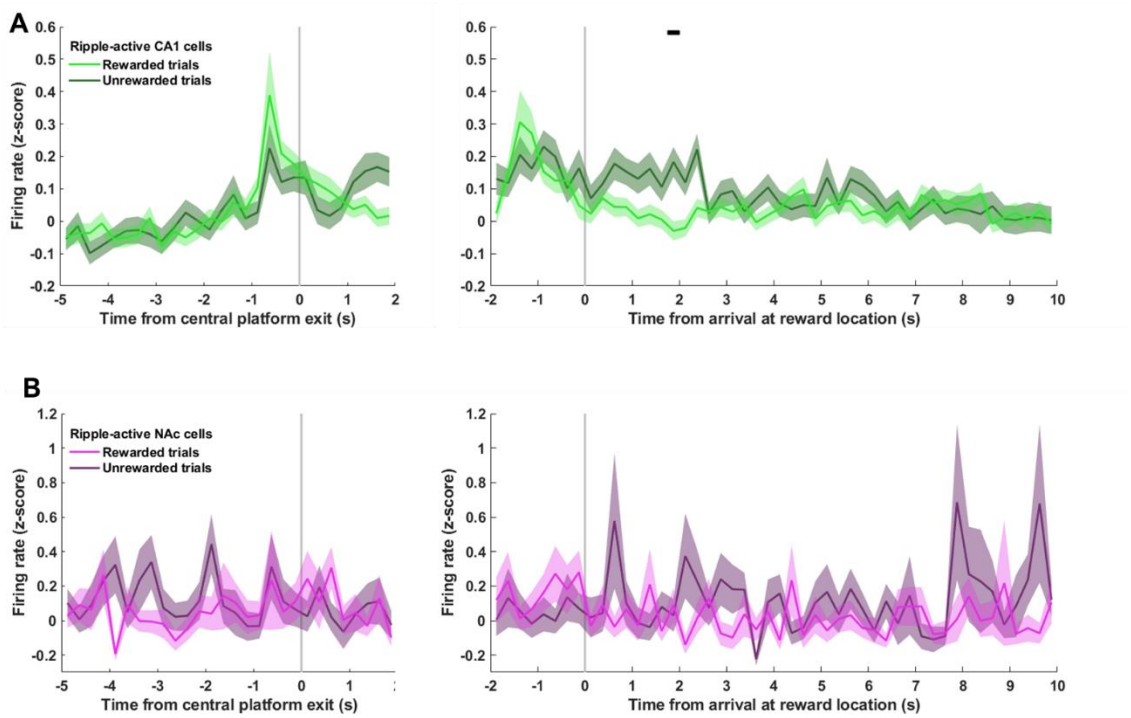

Figure 3.18. Firing rates of ripple-active cells. A. Trial-averaged z-scored firing rates of positively ripple-modulated CA1 cells which fired during a ripple, shown for trials divided into rewarded and unrewarded; black indicates significant difference. B. Trial-averaged z-scored firing rates of positively ripple-modulated NAc cells which fired during a ripple, shown for trials divided into rewarded and unrewarded; no significant differences.

similar, although a low count of trials on which positively modulated cells took part in ripples (n = 173, compared to n = 1866 non-ripple-active) resulted in more inconsistent average firing (fig. 3.17B).

Finally, an additional consideration is the influence of reward outcome on which cells participate in subsequent ripples. Previous studies have found that place cells active during a trial are more likely to be reactivated during ripples at the end of a trial following a reward than no-reward (Singer & Frank, 2009) and more ripples occurred following a rewarded outcome than an unrewarded one (fig. 3.15D). Was there a corresponding association between firing during a trial and participation in reward-associated ripples compared to no-reward-associated ripples? Positively-ripple-modulated cells were restricted to those which took part in ripples during a given trial, and rewarded trials were compared with unrewarded trials. In accumbens, there was no evidence of reward-selectivity, notwithstanding low trial counts (fig. 3.18B). In CA1, a similar peak was observed prior to central platform exit, but there was a slight tendency towards increased firing at the reward location on unrewarded trials (fig. 3.18A). This suggests that reward did not substantially bias ripple activity towards the trajectory just taken, or a particular timepoint of the trial, compared to no-reward.

Taken together, these results show that CA1 and accumbens cells were both modulated by hippocampal ripples, and their participation in ripples during the task was associated with firing at different timepoints during behaviour. As ripples are associated with replay events, a dynamic interaction between task encoding and ripple participation would suggest that replay is directed towards information that is systematically processed by these brain areas at different time points: biased replay. Specifically, cells which increased their firing during TASK ripples also selectively increased their firing rate during the trajectory from central platform to reward location (as well as slightly prior to central platform exit, perhaps reflecting the moment of choice).

3.4. Discussion

Varying cell types, structures, network architectures, neuromodulation and efferent connections make different brain areas suited for distinct kinds of information processing (Sterling & Laughlin, 2015). As a result, different brain areas are involved in diverse aspects of cognition, requiring integration and synchronisation between them to successfully solve cognitive-behavioural problems. Dorsal CA1 and the nucleus accumbens have previously been found to encode spatial information and aspects of reward and uncertainty, respectively, which are both crucial components to learning the task central to this thesis.

These results show that activity in both brain areas is associated with the demands of choosing between navigation options, running towards them, and assessing the rewarding or unrewarding outcome. In the hippocampus, theta, beta and high-gamma activity increased during the trajectory towards reward location, and mid-gamma afterwards. In accumbens there were corresponding increases in theta and mid-gamma activity. And LFP in the two areas showed synchrony in the theta band during trajectory and mid- and high-gamma bands afterwards, indicating that both areas organise their firing during these phases of the task, and communicate information between them.

Sharp-wave ripples, widely associated with replay, memory consolidation and reinforcement learning, occurred after the arrival at the reward location. They occurred at a higher rate following reward, offering a possible mechanism for plasticity during the task by which actions might be reinforced, although evidence was weak for ripple rates increasing following high compared to low reward-prediction error. Accordingly, a large fraction of cells showed firing rate modulation in response to ripples, not only during POST but during TASK as well. The majority of CA1 cells are known to be modulated by hippocampal ripples as excitation spreads throughout the subregion, but the modulation of accumbens cells is more noteworthy because the LFP ripple events themselves are restricted to hippocampus: accumbens ripple-modulation implies functional connectivity between hippocampal cells and accumbens cells. Accumbens cells which were positively modulated by ripples had greater theta modulation (further indicative of their entrainment to hippocampal activity) and a significant peak in firing around the time of central platform entry, coincident with a peak in hippocampal firing. These results open up the possibility of coordinated activity between CA1 and accumbens on this task which is involved in predicting and evaluating action choices, action selection, and movement initiation towards a goal. These possibilities are further explored in subsequent chapters.

Technical problems during this PhD project resulted in a dataset of just one rat with dual-site recordings, which is an unfortunately small sample size and makes it difficult to draw statistically robust conclusions. In addition, a low trial count in each session restricts analyses of, for example, the correlates of accumbens single-unit firing with different combinations of reward location, expectancy, outcome, and prediction error.

Nevertheless, there was a good unit yield and clear time-varying activity which would allow analysis of how activity during behaviour gets replayed during rest and sleep (Chapter 6), to confirm computational hypotheses and predictions about how offline activity can enhance learning from probabilistic rewards (Chapters 4-5).

The involvement of the striatum, in particular, has been the subject of much investigation as a neural implementor of reinforcement learning algorithms which might govern learning in rats, humans and other species (Schönberg et al., 2007; McDannald et al., 2011; Jocham et al., 2011; O'Doherty et al., 2004; van der Meer & Redish, 2010; Li & Daw, 2011; Kim et al., 2009; Ito & Doya, 2011). Computational theory and modelling can reveal the principles that underlie behaviour and suggest the functional significance of neural observations; this is explored in Chapter 4.

Chapter 4: Simulations of reinforcement learning task

4.1. Introduction

The vicarious trial-and-error (VTE) behaviour observed in Chapter 2 has received speculation for decades about what it might reveal about animal cognition and mental representations of the world. Tolman (1948) suggested that VTE may emerge from a cognitive map of the environment, from which rats sequentially generate simulated actions and predict their outcomes as a way to evaluate their options, a theme which has persisted through research in machine learning. A number of algorithms have been developed which are commonly categorised as model-based, creating a model or representation of the world from which simulations can be generated in a style similar to what Tolman (1948) proposed, or model-free, based on simpler and more habitual associations between states or events in the world and corresponding actions. Much of this work in machine learning has proved prophetic for the fields of decision-making and reinforcement learning in animals: the resemblance of phasic dopamine signals to reward-prediction errors is a notable example (Schultz, 1998), but suggestions of replay as a means of speeding up learning pre-dated the discovery of replay in rats (Lin, 1992). Reverse replay, too, was hypothesised as a form of credit assignment before being reported in animals (Cichosz, 1999). The machine learning field therefore has merit for suggesting hypotheses and functions of how replay operates in animals.

Amongst other proposed functions, hippocampal replay has been suggested to aid reinforcement learning, strengthening the associations between states (usually spatial locations or conditioned cues), actions (for example, movement towards another location or a lever press) and outcomes (food reward, electric shock). If these elements are encoded by disparate parts of the brain, offline strengthening of the synapses between them may serve to promote the correct or adaptive behavioural response to the stimulus or

context. The known spatial tuning of hippocampal cells and the reward-responsiveness of accumbens cells (Chapter 3), as well as the observed coordination of activity between them during post-learning rest and sleep (Lansink et al., 2008; Lansink et al., 2009; Sjulson et al., 2018) make these areas obvious candidates for a neural implementation of reinforcement learning online and offline.

4.1.1 Model-free and model-based reinforcement learning

One important consideration is whether the probabilistic task presented in Chapter 2 requires model-free or model-based reinforcement learning. Model-free learning, in particular learning based on temporal-difference (TD) updates, has a long history of use in psychology and neuroscience for modelling both behaviour and neural activity. TD learning is driven by the difference between the estimated future discounted value of a state, action or state-action pair and its observed value, known as reward-prediction error, which closely resembles the responses of dopaminergic midbrain cells. They fire in response to unexpected reward, but after classical conditioning shift their response to a reward-predictive cue and eventually show no response to a predicted reward; instead, firing pauses in response to the unexpected omission of reward (Schultz, 1998). This theory of the function of dopaminergic cells in the ventral tegmental area (VTA) has been useful in describing many aspects of function of the dopaminergic system. Although some discrepancies have been noted between the predictions of TD learning and the observed responses of dopaminergic VTA cells (Redgrave & Gurney, 2006; Gershman & Schoenbaum, 2017) which suggest that VTA is capable of responding to prediction errors more generally, dopamine release (Day et al., 2007) and BOLD responses in the wider dopaminergic system which receive projections from VTA are often consistent with predictions made by model-free reinforcement learning algorithms. Activation of the nucleus accumbens increases in anticipation of reward and in proportion to reward-prediction error (Knutson et al., 2001), showing greater activation in response to unpredictable than predictable reward (Berns et al., 2001), although other studies have attributed this activation to nearby ventral putamen (O'Doherty et al., 2003) or observed the opposite pattern of increased and decreased activation (Pagnoni et al., 2002).

Amongst other extensions of TD learning, the actor/critic framework has been suggested as a possible description of the basal ganglia network (Joel et al., 2002; Sutton & Barto, 1998). In this paradigm, a critic learns to predict rewards and instructs an actor which learns policies for acting in response to states or stimuli, functions which have been ascribed to ventral and dorsal striatum, respectively (O'Doherty et al., 2004; Bornstein & Daw, 2011). As well as correlating with patterns of BOLD activation (O'Doherty et al., 2004; Knutson et al., 2005; Preuschoff et al., 2006), the actor/critic framework is also consistent with lesion studies implicating dorsal striatum in motor control (Burton et al., 2015), and findings that ventral striatal activity correlates better with reward prediction or value prediction than reward-prediction error (Vendrell-

Llopis et al., 2019; Pagnoni et al., 2002; Bissonnette et al., 2013), and perhaps enabling it to convey reward predictions to the VTA to enable the latter to compute reward-prediction errors (Takahashi et al., 2016). A "spiralling" anatomical organisation of connectivity from ventromedial to dorsolateral striatum is also consistent with the principle of the (ventromedial) accumbens conveying a critic signal to the dorsal striatum (Haber et al., 2000; Khamassi & Humphries, 2012).

In more complex environments, learning the value of actions or the association between states and actions is not sufficient for optimal behaviour: transitions between states must also be taken into account, and evidence from devaluation sensitivity and latent learning behaviours suggest that animals do exhibit this kind of cognition (Daw et al., 2005). This is the basis for model-based learning, and appears to evolve in parallel with model-free learning (Yin et al., 2004) to allow flexible responses to task demands. Using this paradigm, the value of an action must take into consideration the likelihood of possible future states, often requiring a computationally costly decision tree to fully predict likely outcomes, but allowing more behavioural flexibility. Inasmuch as VTE can be said to reflect model-based decision-making, the accumbens shows participation in VTE prior to the point of action selection (Stott & Redish, 2014), hinting at an involvement in evaluating outcomes. The accumbens is particularly involved in behaviours requiring flexibility over habitual responses (Nicola, 2010), indicative of model-based learning, and has been suggested to encode rewards associated with particular locations to enable such model-based evaluations (Bornstein & Daw, 2011), perhaps relying on the accumbens shell for model-based learning and core for model-free learning (Bornstein & Daw, 2011). Alternatively, suggestions have been put forward that ventral and dorsal striatum mediate model-based and model-free learning respectively, or that model-based and model-free processes take place in dorsomedial and dorsolateral striatum respectively (Bornstein & Daw, 2011). The convergence of these assorted hypothesised maps of reinforcement learning might lie in a gradient from ventromedial striatum (accumbens shell) encoding value through accumbens core and dorsomedial striatum to dorsolateral striatum encoding habits, reflecting a functional axis of connectivity with other brain regions which allows parallel encoding of several reinforcement learning processes at once (Burton et al., 2015).

Although evidence suggests that model-free and model-based learning systems are both present in animals, they appear to recruit overlapping neural circuits and produce similar behaviour except in carefully constructed experiments (van der Meer & Redish, 2011; Daw et al., 2011), which raises the question of how distinct the processes really are. One suggestion, the dual actor/critic framework, suggests that both the ventral critic and the dorsal actor can be subdivided into model-free and model-based components (Bornstein & Daw, 2011; Khamassi & Humphries, 2012; Colas et al., 2017). But firing in the VTA can reflect model-based reward-prediction errors as well as model-free (Sadacca et al., 2016), contrary to its proposed role in most proposed neural mechanisms for reinforcement learning. Instead, behavioural outputs might depend on a hybrid of both signals (Gläscher et al., 2010; Daw et al., 2011). Moreover, the difference between model-free and model-based learning can be bridged by using model-free learning processes to

update a model-based representation of the world: offline replay has been proposed as a way to achieve this (van Seijen & Sutton, 2015).

Dyna algorithms (Sutton & Barto, 1998) are an extension of model-free learning, but create a hybrid between the two taxonomies. They maintain a model of the transitions between actions and states, but rather than computing the expected cumulative future reward over multiple successive actions, the value of each state or action is cached by back-propagating values from a reward location to each predecessor, effectively deriving long-term values iteratively and storing them in a simple lookup table to allow fast action selection. Updating values for each state or action involves much simpler, model-free processes, but performing them in sequence allows the whole cognitive map to be updated at a smaller computational cost (Sutton & Barto, 1998; Russek et al., 2017). This has been conceptually linked to the observation of reverse replay which is observed at a particularly high rate immediately following reward: the suggestion is that reverse replay might serve as a form of credit assignment to propagate the newly observed value of the animal's current state backwards to previous place fields (Johnson & Redish, 2005; Mattar & Daw, 2018). A related idea is successor representation (Gershman et al., 2012), in which the expected future occupancy of spatial states is represented, and values are propagated through the cognitive map from one state to its successor according to the expected future occupancy: this serves effectively as a more efficient form of tree search through future possibilities. The combination of these methods – successor-representation-Dyna, or SR-Dyna – has been shown to achieve learning behaviours that model-free algorithms fail to replicate, including latent learning and reward revaluation (Russek et al., 2017).

The observation that disrupting sharp-wave ripples (and therefore replay) during awake behaviour disrupts learning can be taken as evidence that awake replay serves to propagate reward information through the cognitive map (Jadhav et al., 2012). Many other observations about how awake replay interacts with the cognitive demands of learning have been replicated using a framework in which experiences are prioritised for replay by a balance between the likelihood that an experience will be encountered again in the near future, and its usefulness for updating existing information (determined by reward-prediction errors; Mattar & Daw, 2018). These theories offer hypotheses for how replay might support reinforcement learning on spatial tasks.

4.1.2 Replay in machine learning

Efforts in machine learning have attempted to replicate, or take inspiration from, biological replay to make learning more efficient in various artificial applications, including for robotics (Adam et al., 2012) and game-playing (Wang et al., 2016). In theory, storing experiences in a memory buffer and sampling from them offline can make more efficient use of limited experience, improving the sample efficiency to speed up

learning with respect to the amount of online learning required (Kalyanakrishnan & Stone, 2007), but in practice machine learning research has found that the computational costs of storing experiences, calculating which ones to sample, and performing the offline updates can outweigh the costs of simply sampling more from the environment (van Seijen & Sutton, 2013; Liu & Zou, 2018). Replay is therefore most useful if exposure to the environment is limited but time is abundant, which is usually the case for animals which spend some limited time foraging and a lot of time resting: it may be more beneficial to rest and replay experiences from memory than continue foraging to gain more real-world experience (especially when foraging risks injury, predator attacks and other dangers). Additionally, there are certain advantages to interleaving new and old, or ongoing and recent, experiences: in machine learning this is described as breaking temporal correlations (Mnih et al., 2013, 2015; Schaul et al., 2016) but in cognitive science it can be thought of as generalising between episodes for acquiring latent knowledge (Kumaran et al., 2016).

Often, experience replay in machine learning samples uniformly from the memory buffer (Mnih et al., 2015; Tessler et al., 2016), but more can be achieved by biasing replay in adaptive ways. For example, flexible learning can be achieved by biasing replay which allows the statistics of the perceived environment to be skewed to match the agent's goals: virtually experiencing reward locations allows more to be learned from goal locations (Lin, 1992; Kumaran et al., 2016). Replaying successful, i.e. rewarded, information can selectively reinforce actions which are beneficial (Narasimhan et al., 2015; Isele & Cosgun, 2018), but this is applicable only for situations with deterministic rewards where there is a straightforward relationship between "good" actions and beneficial outcomes. Using reward-prediction errors to prioritise which experiences are replayed can increase learning speed further in a range of tasks which involve sparse but generally deterministic rewards (Moore & Atkeson, 1993; Schaul et al., 2015; van Seijen & Sutton, 2013), but taken to an extreme it can lead to catastrophic forgetting of known, unsurprising information (Isele & Gosgun, 2018). Further possibilities are that more uncertain experiences may be played more (to reduce uncertainty), or rarer experiences may be replayed more (to counteract the lack of real-word experience), if the experiences in question may be motivationally relevant (Isele & Cosgun, 2018).

These studies demonstrate that the costs and benefits of different replay policies are task-dependent, and research examining replay for learning from stochastic rewards are lacking, in both the neurophysiological and machine learning fields. To test how replay might affect learning in the probabilistic maze task, simulations of the task were run with a reinforcement learning algorithm, and various policies for replay were applied to observe the effect on speed and accuracy of learning. A replay policy which enhances learning in the simulation might also be implemented by rats performing the same task, so this is instructive for forming hypotheses about biological replay.

In order to test ideas about replay, an appropriate reinforcement learning algorithm was first considered. Q-learning (described below) was selected because of its simplicity and broad applicability to many

reinforcement learning scenarios, and because of previous work which has found Q-learning-like neural correlates (Ito & Doya, 2009) and similarities to hippocampal replay (Aubin et al., 2018).

4.1.3 Q-learning

Q-learning is a common temporal-difference-based learning algorithm for Markov decision processes. In Q-learning, an agent selects actions in its environment and observes the outcome, recording at each time step t its starting state s_t, selected action a_t, resulting reward r_t, and resulting state s_{t+1}. In a maze task, states may correspond to locations in the environment, and/or cues that indicate task-relevant information such as where a reward might be found, and actions may be moves in a particular direction. The agent's goal is to build up estimates of Q-values for every state (or for every state-action pair) corresponding to the future discounted expected reward, which is the temporal difference between the current state and the reward state. These Q-value estimates can then be used to guide actions to maximise reward. In the first instance actions are selected randomly, but at each time step the Q-value for the state-action pair observed can be updated:

$$Q(s_t, a_t) \leftarrow (1 - \alpha) \cdot Q(s_t, a_t) + \alpha(r_t \cdot \gamma \cdot \max Q(s_{t+1}, a))$$

where $\alpha \in (0,1)$ is a learning rate parameter which determines the degree to which new information overrides old information, and $\gamma \in (0,1)$ is a discount parameter which determines the importance of long-term gains. Q-learning has been used to model biological maze learning, and has been proposed as a model for cortico-striatal plasticity. Reward-prediction errors, the difference between expected reward and actual reward which drives the update of Q-values, resembles quite closely the dopaminergic input of VTA to striatum, and the fact that dopamine drives cortico-synaptic plasticity in the striatum is further evidence for the suitability of Q-learning as a model of mammalian maze learning. Watkins & Dayan (1992) showed that Q-learning converges to the optimum action-values, provided that the agent is given sufficient experience of all state-action pairs. Q-learning is particularly useful for sequential learning in environments where rewards are sparse: this is the case if a maze environment is divided into many states, which can represent place fields, such that the agent must pass through many place fields to reach a reward location. For this Chapter, however, a simpler construction of the task was considered, in which a state covers a whole arm of the maze, and therefore every state may elicit a reward. Q-learning was combined with replay following each online experience: for every trial Q-values were updated and the experience's tuple $\{s, a, r, s'\}$ was added to a memory buffer, and k samples were selected from the memory buffer (usually at random) and used to further update Q-values. This has the effect of learning from several trials per actual trial of experience. Updating Q-values based on both online and offline experience has been shown to speed up learning, compared to traditional online Q-learning alone (Sutton, 1990).

4.1.4 Aims of this chapter

1. To determine whether Q-learning is capable of solving the stochastic reinforcement learning problem presented in Chapters 2 and 3
2. To qualitatively assess the suitability of Q-learning for modelling the behaviour presented in Chapter 2

4.2. Methods

Q-learning. A Q-learning algorithm was trained on simulations of the task. The environment was defined by three states, $s = \{1,2,3\}$, and three actions, $a = \{1,2,3\}$, which corresponded to the arms of the maze. The state was the arm visited on the previous trial, such that for trial t, $s_t = a_{t-1}$. s_1 was set arbitrarily to 1. Rewards $r_t = \{0,1\}$ were delivered probabilistically as described in the Results.

Q-values $Q(s_i, a_i)$ were maintained for all 9 state-action pairs, initialised to a negligible value above 0 (no expected value; negligible value added to prevent computation errors) and updated on each trial t according to the trial's tuple $\{s_t, a_t, r_t, s'_t\}$:

$$Q(s_t, a_t) \leftarrow (1 - \alpha) \cdot Q(s_t, a_t) + \alpha(r_t + \gamma \cdot \max Q(s_{t+1}, a))$$

Values of parameters α and γ are described in the Results. Except where otherwise stated, the same simulation was run 1,000 times and averaged to overcome randomness in rewards and action selection.

Action selection. Actions were selected probabilistically according to:

$$p_{s,a_i} = \left(\frac{Q(s, a_i)}{\sum_{i=1}^{3} Q(s, a_i)} + \frac{\varepsilon}{3}\right)(1 + \varepsilon)^{-1}$$

When $\varepsilon = 0$, action probabilities are directly proportional to Q-values; with increasing values of ε the action probabilities become more equalised and therefore closer to random action selection. Example action probabilities are illustrated for a set of Q-values [0.1, 0.4, 0.6] with different values of ε (fig. 4.1).

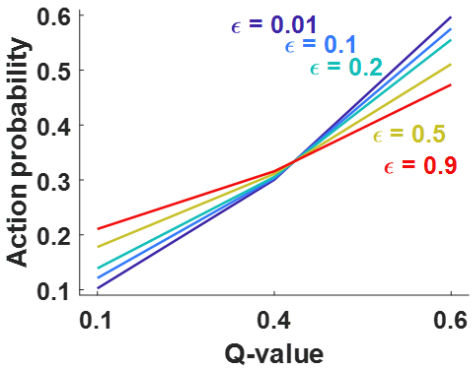

Figure 4.1. Action probabilities. Action probabilities for a set of Q-values [0.1, 0.4, 0.6] corresponding to three available actions in a given state. Five sets of action probabilities calculated with different values of ε are shown.

Replay. In simulations which incorporated replay, the tuple for every trial was stored in a memory buffer. After every 50th online trial, a number of samples as described in the Results were selected for replay. Samples were chosen probabilistically according to different replay policies (described below), and current Q-values were updated with the sampled tuple.

Pure random replay policy. All trials experienced n_t up to the current trial t were selected with probability $p_i = \frac{1}{n_t}$.

Recency-biased replay. Trials were sorted according to their recency, i, and were selected according to a specified recency parameter φ, with probability $p_i = i^{\varphi}$.

Rewarded-only replay. The subset of trials experienced which were not rewarded were discarded, and the remaining (rewarded) trials were selected according to the same specified recency parameter φ, with probability $p_i = i^{\varphi}$.

RPE-biased replay. For every trial, the reward-prediction error rpe_t was calculated as

$$\mathrm{rpe}_t = (1-\alpha)\cdot Q(s_t, a_t) + \alpha(r_t + \gamma \cdot \max Q(s_{t+1}, a)) - Q(s_t, a_t)$$

i.e. the difference between the updated Q-value and the old Q-value. At the point of replay, all trials experienced up to the current trial were ranked according to the magnitude of their RPE, and were selected with probability $p_i = i^{\varphi}$.

Rewarded-state-action-pair replay. First the state-action pair s, a to be sampled from was chosen probabilistically according to its Q-value, $P(s, a_i) = (\frac{Q(s,a_i)}{\sum_{i=1}^{3} Q(s,a_i)} + \frac{\varepsilon}{3})(1+\varepsilon)^{-1}$. The subset of trials experienced of the chosen state-action pair were then sorted according to their recency, i, and were selected according to a specified recency parameter φ, with probability $p_i = i^{\varphi}$.

RPE state-action-pair replay. For each state-action pair s, a, a weighted absolute sum $\mathrm{wrpe}_{s,a}$ was calculated of the RPE of all trials experienced of this pair, with trials weighted by their recency i raised to the power of the recency factor φ:

$$\mathrm{wrpe}_{s,a} = \left| \sum_{i=1}^{n} \mathrm{rpe}_i \cdot i^{\varphi} \right|$$

One state-action pair was then chosen for replay according to the probability $P_{s,a} = \frac{\mathrm{wrpe}_{s,a}}{\sum_{s=1}^{3}\sum_{a=1}^{3}\mathrm{wrpe}_{s,a}}$, and from this state-action pair the subset of trials experienced of it were sorted according to their recency, i, and were selected according to a specified recency parameter φ, with probability $p_i = i^{\varphi}$.

4.3. Results

Before examining the effect of different replay policies with Q-learning, a baseline performance was established and examined for Q-learning with no replay.

4.3.1. Q-learning in a stationary environment

First, a Q-learning algorithm was trained to navigate the three-armed maze with static reward probabilities of [0.75, 0.5, 0.25]. Initially, the parameter values were chosen as $\alpha = 0.1$, $\gamma = 0.01$, $\varepsilon = 0.3$. An example run of 100 trials is shown (fig. 4.2A), over which the Q-learning agent obtained 56 rewards (fig. 4.2B), substantially more than the 33 rewards expected by chance. This was achieved by performing optimal actions at a higher rate than chance (fig. 4.2B), i.e. alternating between the high- and mid-probability arms. The Q-values for state-action pairs which drove the action selection (fig. 4.2E) converged to values close to the average reward obtained for each state-action pair (fig. 4.2D).

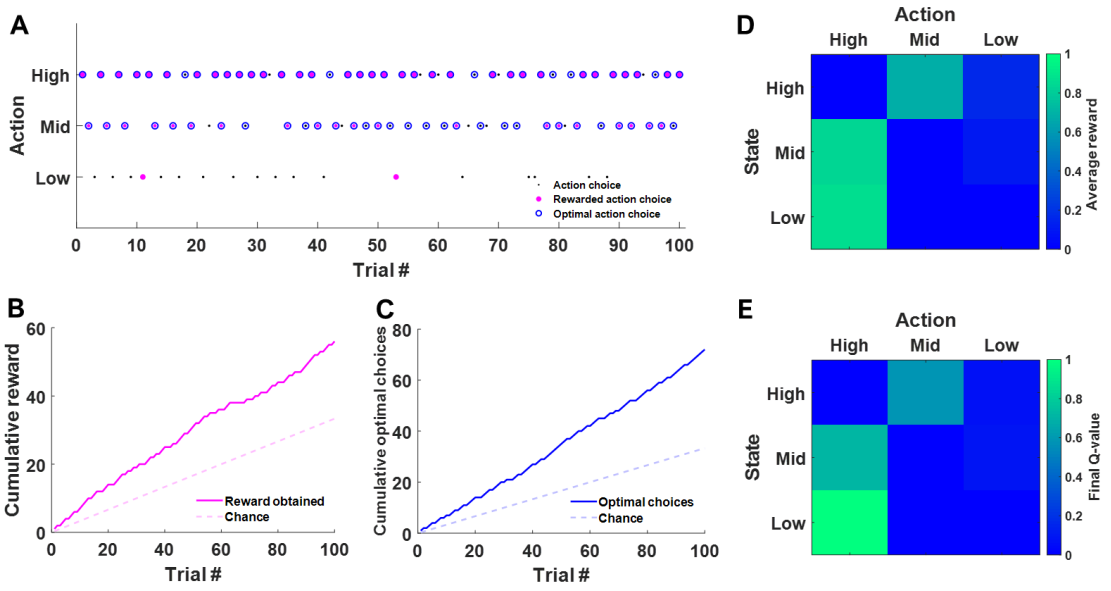

Figure 4.2. Example Q-learning simulation run of 100 trials. **A.** Action (arm) chosen on each trial. Optimal actions (the high arm, or the mid arm when high on the previous trial) are indicated with blue circles. Rewards were probabilistic at all arms; rewarded trials are in pink. **B.** Cumulative reward obtained on the run. **C.** Cumulative optimal choices made on the run. **D.** Average reward obtained for each transition from one arm on one trial (the state) to the arm on the subsequent trial (action). Revisits to the arm visited on the previous trial, i.e. where action = state, were never rewarded. **E.** Q-values for all state-action pairs at the end of 100 trials, updated throughout learning according to rewards obtained in D.

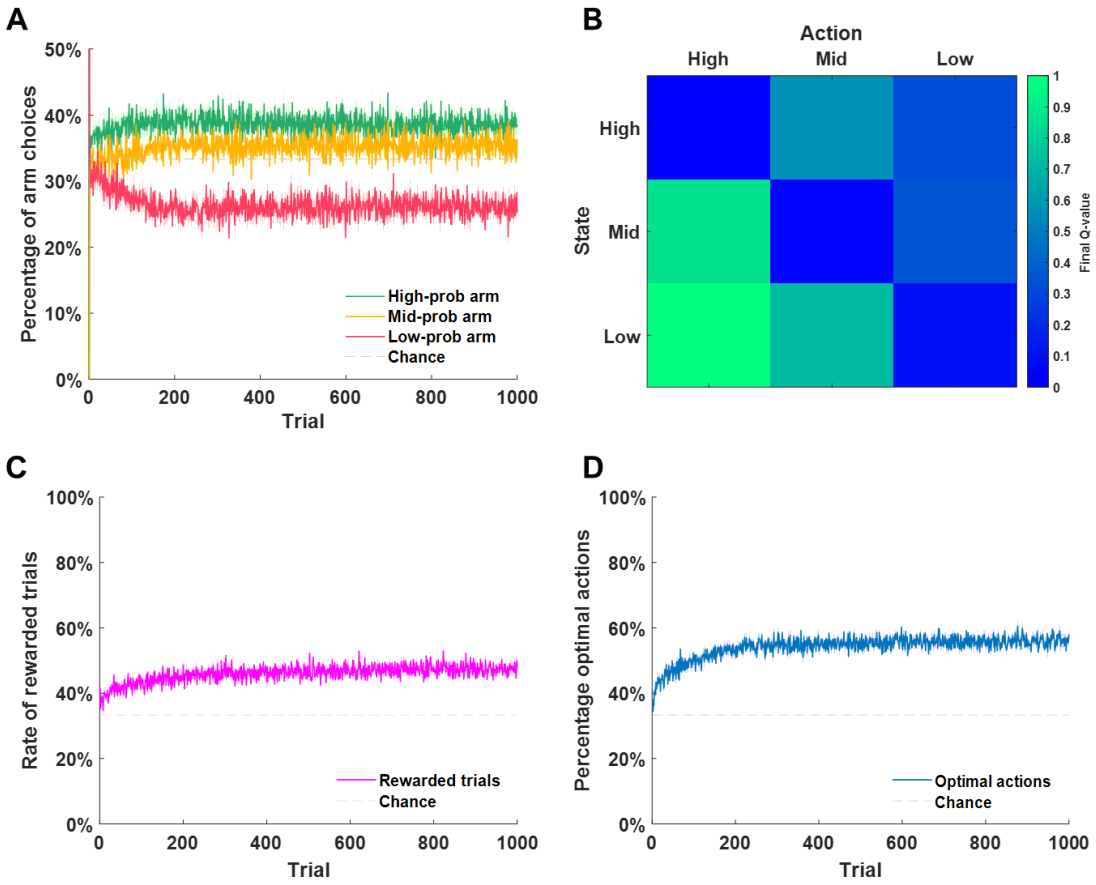

Figure 4.3. Model performance averaged over 1,000 runs. **A.** Rate of entry to each arm. **B.** Q-values for state-action pairs at the end of learning. **C.** Rate of obtaining rewards. **D.** Rate of optimal performance. In A, C and D data shown are a moving average of 10 trials over 1,000 runs, for better visualisation.

The simulation was repeated 1,000-fold to obtain a measure of average performance, and extended to 1,000 trials to see the stability of behaviour. The increase in performance throughout training was apparent, in terms of both rewards obtained (fig. 4.3C) and optimality of choices (fig. 4.3D), but only up to about 300 trials after which optimality plateaued at approximately 56%. The rate of optimal behaviour remained well below the theoretical maximum of 100%, and the rate of obtaining rewards well below the theoretical maximum average of 62.5%. Importantly, with these parameters the agent exhibited probability-matching (fig. 4.3A), selecting each arm in a similar proportion to its reward probability, replicating the same behaviour in rats (Chapter 2).

Next, the parameter state-space was searched to find the optimal parameter values for learning this stationary task. This was done iteratively, such that first the value of the learning rate α was varied between 0 and 1, keeping discount factor γ at 0.01 and exploration factor ε at 0.3. Predictably, with a learning rate of 0.0 performance remained at chance because no learning took place (fig. 4.4A). The best performance was reached with a α value of 0.7 (64% of actions optimal over trials 501-1000; fig. 4.4A), so subsequently the γ value was varied between 0 and 1, with α at 0.7 and ε at 0.3. The best performance was reached

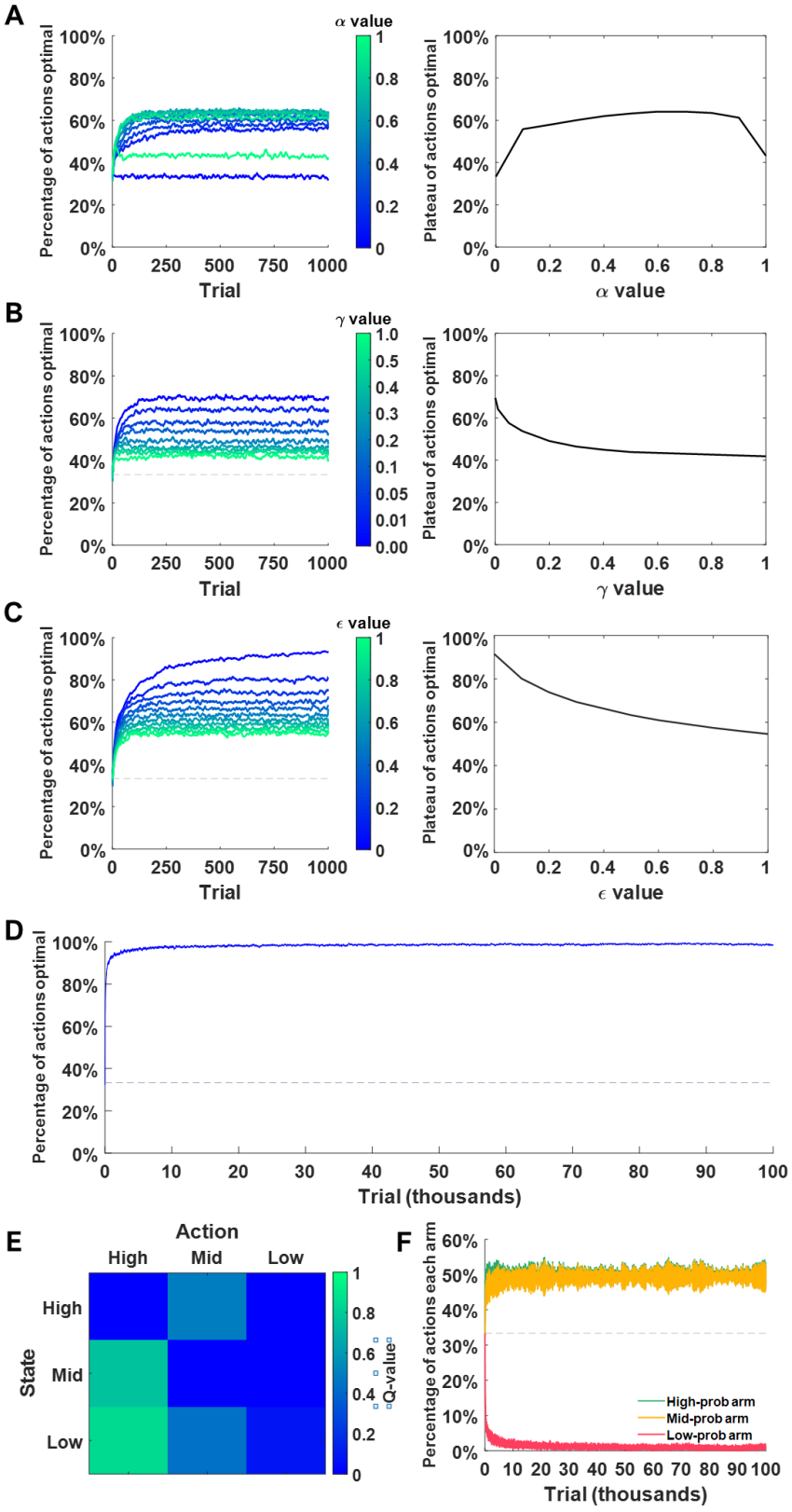

Figure 4.4. Parameter perturbations. **A-C.** Rate of optimal action selection, average over 1,000 runs, for varying values of α (A), γ (B), and ε (C). Left: optimal behaviour over all trials. Right: plateau of optimal behaviour, i.e. average over trials 501-1,000. **D-F.** Performance with near-optimal parameter values over 100,000 trials. **D.** Rate of optimal behaviour. **E.** Q-values at the end of learning. **F.** Rate of choosing each arm. See main text for parameter values.

with a γ value of 0.00 (69% of actions optimal over trials 501-1000; fig. 4.4B), so finally the ε value was varied between 0 and 1, with α at 0.7 and γ at 0. The best performance after that was with a ε value of 0 (92% of actions optimal over trials 501-1000; fig. 4.4C). In fact, at these latter parameter values the performance did not reach a plateau within 1000 trials, so the simulation was re-run over 100,000 trials. Over the extended training, the model converged at 98% optimal behaviour over trials 99,501-100,000 (fig. 4.4D) with almost no probability-matching (fig. 4.4F). This demonstrates that the Q-learning algorithm is capable of learning this simple task to near-perfection, given enough trials and the right parameter values, but that different rates of learning and maximum performance can be captured by varying the parameter values.

4.3.2 Q-learning in a non-stationary environment

So far, Q-learning was trained only in a static environment, meaning that the reward values did not change and therefore the Q-values and behavioural policy did not need to change. To test the capability of Q-learning for flexibility in a dynamic, non-stationary environment, the reward probabilities were changed every 500 trials according to the procedure used in Chapter 2. First the high-probability arm was changed from 75% to 87.5% and the low-probability arm was changed from 25% to 12.5% (the mid-probability arm remained unchanged). Then the high- and low-probability arms were switched. An extremely low rate of exploration ($\varepsilon = 0$) would no longer be as advantageous in this environment because it would take longer to discover that better reward probabilities are available at different state-action pairs.

Reverting to the original parameter values used in the stationary environment (learning rate $\alpha = 0.1$, discount factor $\gamma = 0.01$, exploration rate $\varepsilon = 0.3$), the agent learned to adapt its behaviour to the

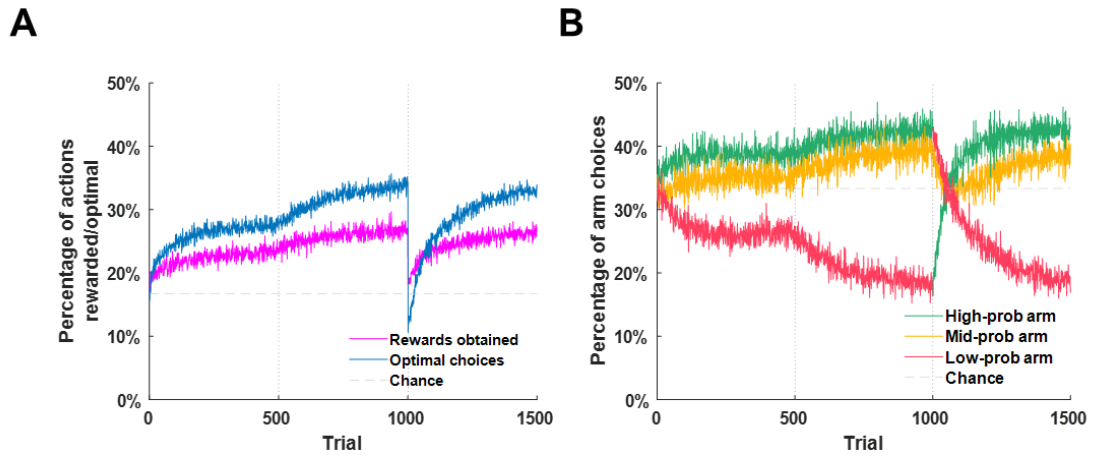

Figure 4.5. Model performance in a non-stationary environment. A. Rate of optimal behaviour and resulting rewards obtained in a non-stationary environment with changing reward probabilities. B. Rate of selecting each action. Vertical dashed lines indicate when reward probabilities were changed.

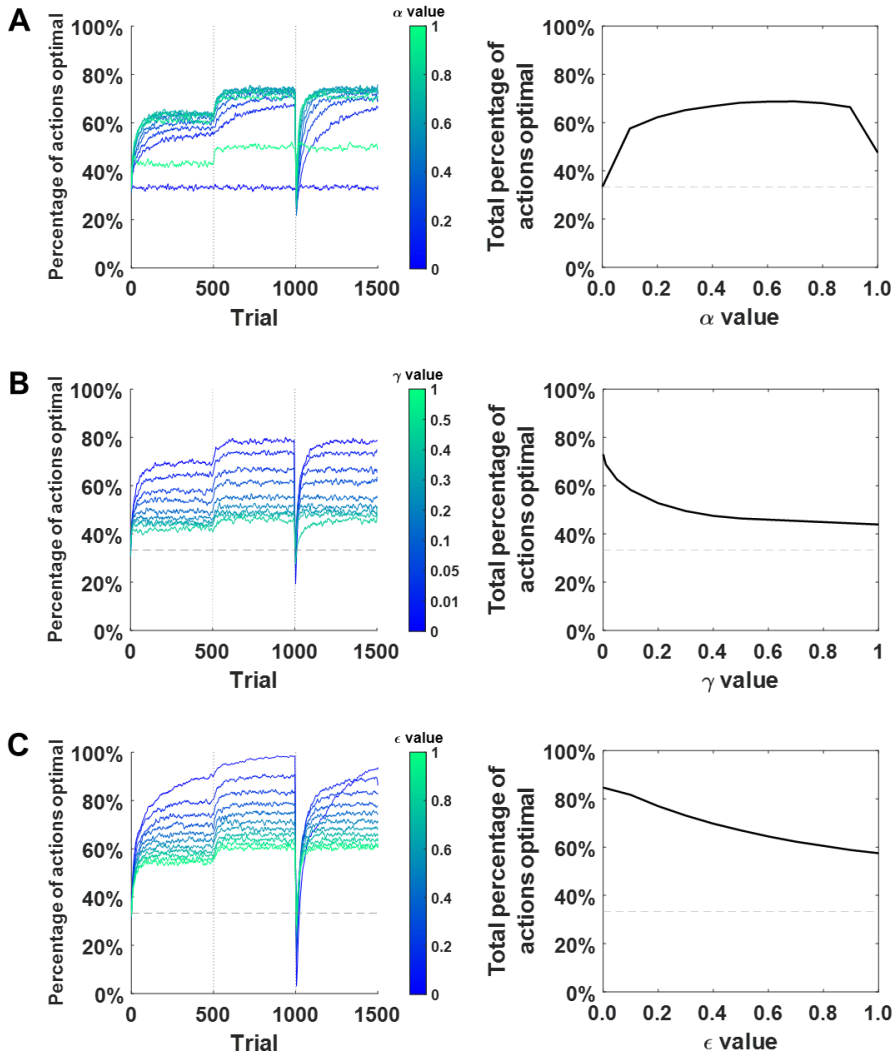

Figure 4.6. Parameter perturbations. Rate of optimal action selection, averaged over 1,000 runs, for varying values of α (**A**), γ (**B**), and ε (**C**). Vertical dashed lines indicate when reward probabilities changed. Left: optimal behaviour over all trials. Right: overall rate of optimal behaviour, i.e. average over all trials.

changing probabilities (fig. 4.5). Under the first set of reward probabilities (identical to the stationary environment), it reached a plateau of 55% over trials 401-500. Under the second set, rate of action selection changed and optimal performance increased to a plateau of 66% over trials 901-1000. Under the third set, behaviour rapidly changed in response to vastly different reward probabilities and reached a similar plateau to the second set of reward probabilities: 66% over trials 1401-1500. The Q-learning algorithm, therefore, can show flexibility in the face of changing task demands.

As in the non-stationary environment, parameter values were systematically varied to observe the effect on learning. First α was varied between 0 and 1, keeping γ at 0.01 and ε at 0.3. The best overall rate of optimal performance was found at a α value of 0.7, as for the non-stationary environment, at which 69% of all actions were optimal (fig. 4.6A); at this learning rate, higher peak performance was reached, and it was

reached in fewer trials. Next, γ was varied between 0 and 1, with α at 0.7 and ε at 0.3. The best performance was reached with a γ value of 0.00, at which 73% of all actions were optimal (fig. 4.6B). The best performance was reached with a ε value of 0.0, at which 85% of all actions were optimal (fig. 4.6C), but it is notable that the re-learning of the behavioural policy in trials 1001-1500 was markedly slower: the agent was slower to adapt to the changing environment.

4.3.3 Q-learning with replay

Replay has been used with Q-learning models to enhance learning in a variety of tasks. To investigate the possible roles of replay in learning on this task, simulations were run as previously but with a "sleep session" every 50 trials. All trials were stored in a memory buffer and, at these points, samples were selected from the memory buffer to be replayed and perform updates to the Q-values accordingly. Six policies by which samples could be chosen were compared: pure random replay, recency-biased replay, rewarded-only replay, RPE-biased replay, rewarded-state-action replay, and RPE-state-action replay (see Methods for details).

Pure random replay

Under the policy of pure random replay, every 50 online trials a number of samples were selected from the memory buffer to be replayed, selected at random from the total experience so far. Varying the number of replays showed that, in general, more replay equated to worse performance compared to no-replay Q-learning performance (fig. 4.7A; no replay shown in black markers). Specifically, there was a sharp drop in performance every 50 trials, corresponding to when the replay events took place, before the performance slowly increased again following more online trials. When the reward probabilities changed, even the peak performance prior to replay was below the plateau achieved under no replay: offline replay events outweighed online replay events and diverted the Q-values away from optimum.

This result is unsurprising, because under the pure random replay policy early samples were just as likely to be replayed and used to update Q-values as later samples, even when early samples represent outdated information about the state of the environment. The effect was to bias Q-values towards what they used to be, undermining Q-learning's flexibility and ability to adapt to new reward conditions. Interestingly, however, the first session of replay events after trial 50 led to a sudden increase in performance above the no-replay baseline: at this point, with static reward probabilities, it accelerated learning.

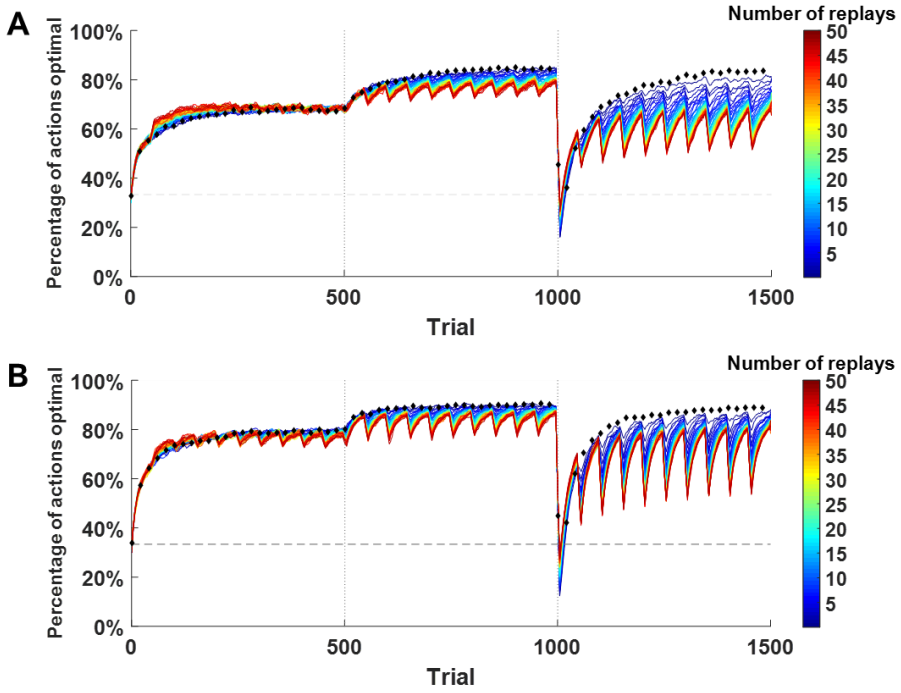

Figure 4.7. Pure random replay. Rate of optimal action selection, averaged over 1,000 runs, under a policy of pure random replay. Samples were replayed from the memory buffer after every 50th trial. A. Performance using near-optimal parameter values, with varying numbers of replay updates per session. B. Performance with a suboptimal learning rate, with varying numbers of replay updates per session. Black diamond markers indicate performance at no-replay baseline.

This suggests that with a lower and less optimal learning rate, random replay might serve to speed up initial learning. The α value was therefore lowered to 0.3, a value previously shown to decrease performance, for comparison. Initial learning from trial 50 was improved by random replay with this lower learning rate (fig. 4.7B): although peak performance was not increased, the plateau was reached sooner. This benefit of replay disappeared when the reward probabilities changed, unsurprisingly, because outdated samples were replayed as often as up-to-date ones.

Random replay, therefore, can speed up learning when the learning rate is suboptimal, but undermines the flexibility of Q-learning when it comes to adapting to the changing environment. A straightforward solution is to bias the samples being replayed by how recently they were experienced, such that recent samples get replayed more than older samples; this was the approach taken in the recency-biased replay policy.

Recency-biased replay

As with random replay, after every 50 online trials a number of samples were selected from memory to be replayed. The probability of each sample being selected was scaled by a power function according to its

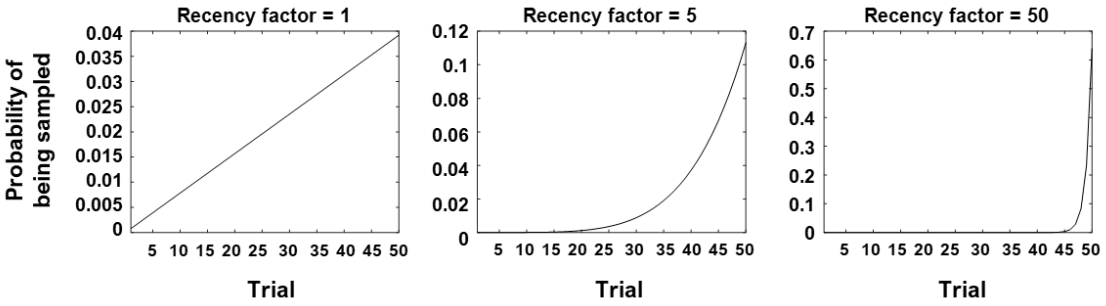

Figure 4.8. Recency factor. Recency factor changed the skew towards more recent samples over 50 trials.

recency, such that more recent trials were more likely to be replayed (fig. 4.8; see Methods). Both the recency factor (i.e. how strong the weighting was towards more recent trials) and the number of replays could be varied.

First, the recency factor was varied between 1 and 50, with parameters $\alpha = 0.7$, $\gamma = 0.0$, $\varepsilon = 0.1$, and the number of samples to be replayed at each interval of 50 offline trials set to 5. This form of replay made little difference to performance under the first two reward conditions, but made a big difference for reversal learning (fig. 4.9A). With low recency factors, performance dropped dramatically following each set of replays. This is because older samples from the first two reward conditions were often selected, despite offering out-of-date information about the environment, which skewed the Q-values back towards the old reward conditions. Under higher recency factors, this effect was diminished because older samples were much less likely to be replayed. Nevertheless, even under high recency factors, performance was no better than the baseline of no-replay with Q-learning (fig. 4.9A, shown in black markers). This was apparent when the number of replays was varied between 1 and 50, keeping the other parameters at $\alpha = 0.7$, $\gamma = 0.0$, $\varepsilon = 0.1$ and recency factor 50. The more replays, the slower the agent was to reach the plateau of peak performance, both for the initial learning and the reversal learning, resulting in poorer overall performance (fig. 4.9B). The detrimental effect of replay on reversal learning is a result of replaying outdated information, as discussed; the detrimental effect on initial learning is more interesting because none of the information is out of date. Instead, replay here may be amplifying the effect of early trials which are necessarily unrepresentative of the environment statistics.

So far recency-biased replay appears only to worsen performance, but this might be because performance was already close to maximum: the α, γ and ε values were chosen specifically to maximise performance for Q-learning. Recency-biased replay might be more beneficial when other learning parameters are suboptimal. For this reason, the number of replays was varied again between 1 and 50, with α lowered to 0.3. Under these parameter values, recency-biased replay did improve performance, reaching higher peak performance than the Q-learning baseline (fig. 4.9C). This was true for all three reward conditions, demonstrating that recency-biased replay enhances rather than undermines flexibility.

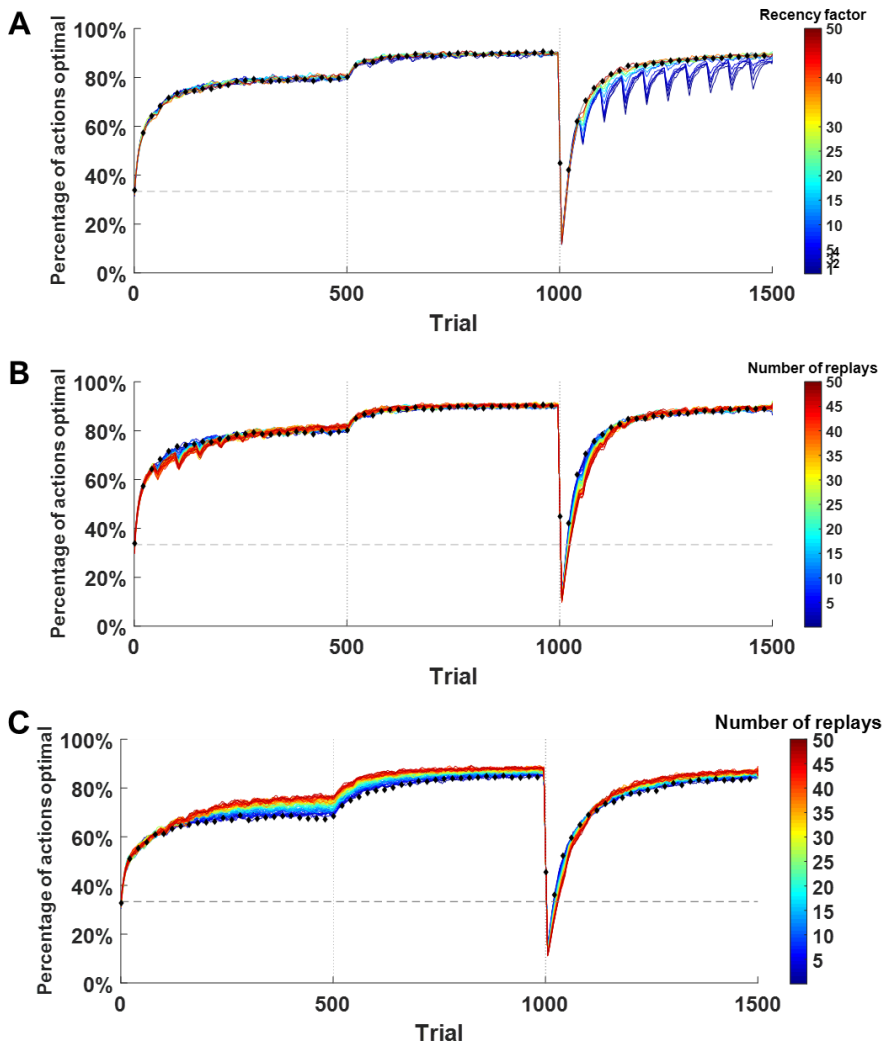

Figure 4.9. Recency-biased replay. Rate of optimal action selection under a policy of recency-biased replay. **A.** Performance using near-optimal parameter values, with varying recency factor for skewing replay samples towards more recent trials. **B.** Performance using near-optimal parameter values, with varying numbers of replay updates per session. **C.** Performance with a suboptimal learning rate, with varying numbers of replay updates per session. Black diamond markers indicate performance at no-replay baseline.

Rewarded-only replay

One suggestion for deterministic tasks has been to replay selectively the state-action pairs which lead to successful outcomes, i.e. to replay rewarded trials only. To test this, replay was performed biased by recency, as above, but with the pool of possible samples limited to rewarded trials. Suboptimal parameter values of $\alpha = 0.7$, $\gamma = 0.0$ and $\varepsilon = 0.1$ were used. Varying the recency factor (fig. 4.10A) and number of samples replayed (fig. 4.10B) showed no improvement in performance, and a decrease in performance with higher replays or lower recency factors. This is because replaying only rewarded trials provides a

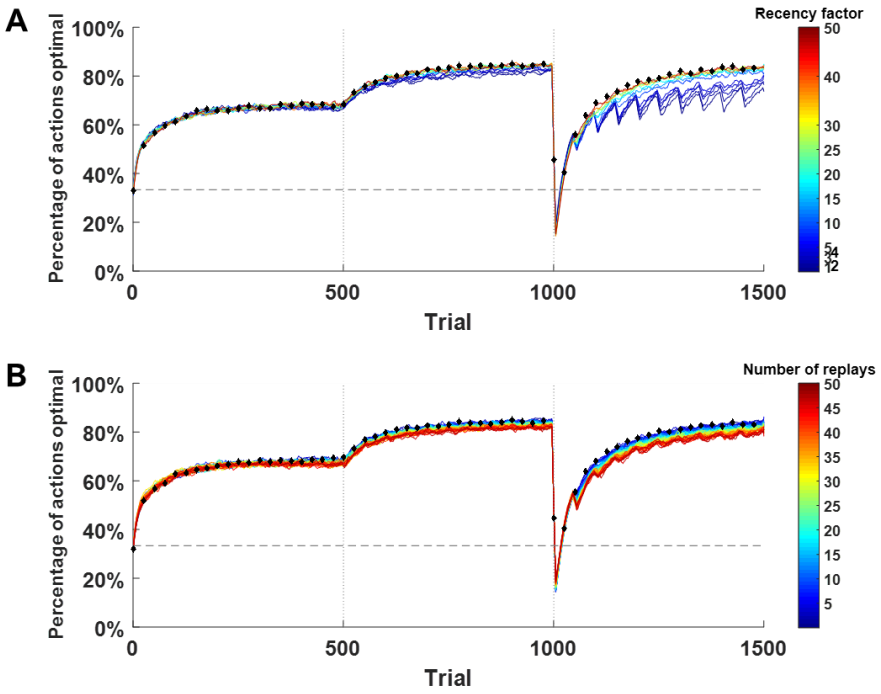

Figure 4.10. Rewarded-only replay. Rate of optimal action selection under a policy of rewarded-only replay. **A.** Performance with varying recency factor for skewing replay samples towards more recent trials. **B.** Performance with varying numbers of replay updates per session. Black diamond markers indicate performance at no-replay baseline.

training set of offline learning which is unrepresentative of the environment: sometimes the same action performed in the same state leads to no reward, which is not reflected in the replayed samples. Crucial to this task is acquiring the reward probabilities associated with each state-action pair, which is lost if only rewarded trials are replayed.

RPE-biased relay

An alternative replay policy is to prioritise learning from the most surprising or unexpected outcomes, because a prediction error indicates that the internal model is incorrect. Accordingly, a replay policy was implemented in which trials were prioritised by the reward-prediction error (difference between expected value and observed value; see Methods) they elicited. Under this policy, performance was similarly worse than with no replay (fig. 4.11), because selecting samples in this way also presents an unrepresentative distribution of the environment to the model during replay sessions, similarly to the problem with rewarded-only replay. If, for example, a no-reward outcome after transitioning from the mid-probability arm to the high-probability arm (which usually results in a reward) elicits a high reward-prediction error, this will be

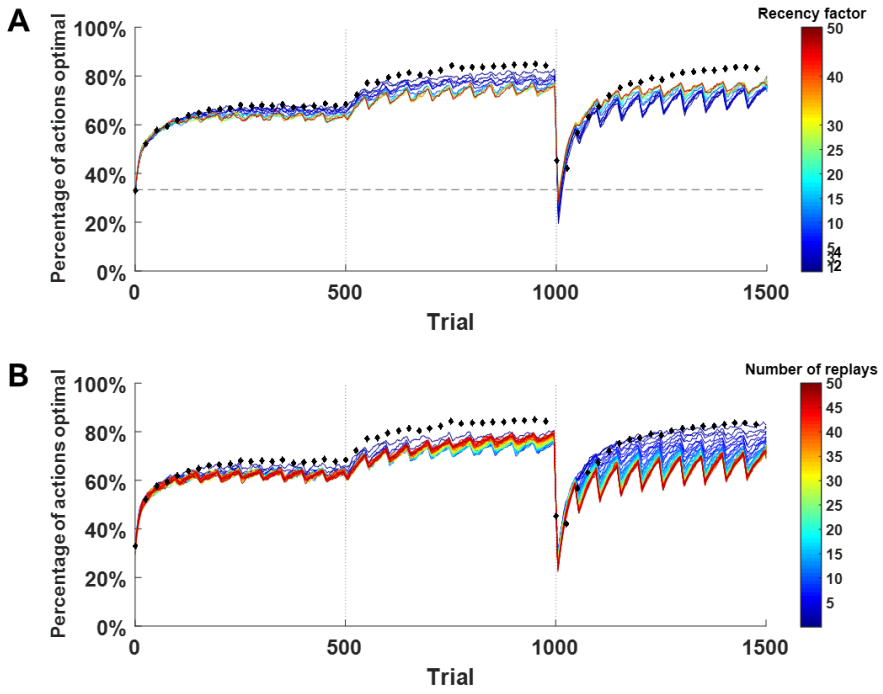

Figure 4.11. RPE-biased replay. Rate of optimal action selection under a policy of RPE-biased replay. **A.** Performance with varying recency factor for skewing replay samples towards more recent trials. **B.** Performance with varying numbers of replay updates per session. Black diamond markers indicate performance at no-replay baseline.

prioritised for replay; but over-representing this type of trial in the training leads to an undervaluing of this state-action pair.

More generally, the problem with this kind of replay policy is that in a stochastic environment every individual trial is unrepresentative, because reward outcomes are binary but reward probabilities are not. Prioritising individual types of trial is therefore inappropriate. Instead, pooling together all (recent) trials from one state-action pair can give a more accurate representation of the distribution of rewards from state-action pairs; this method of prioritisation is considered next.

Rewarded-state-action-pair replay

Selectively replaying rewarded trials lead to a decrement in performance, but prioritising state-action pairs according to their average reward may offer a better way to bootstrap learning without skewing it towards unrepresentative trials. Therefore, for every replay, the average recent reward obtained for each state-action pair was approximated by taking its Q-value at the current time; one state-action pair was selected in proportion to its Q-value (see Methods), and from the pool of exemplar trials of this state-action pair, one

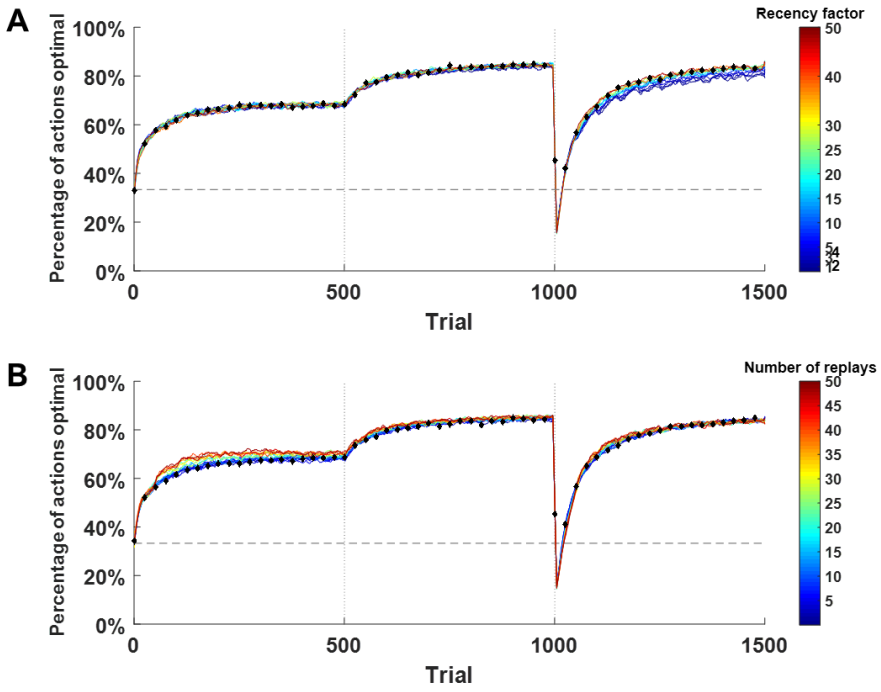

Figure 4.12. Rewarded state-action-pair replay. Rate of optimal action selection under a policy of rewarded-state-action-pair replay. **A.** Performance with varying recency factor for skewing replay samples towards more recent trials. **B.** Performance with varying numbers of replay updates per session. Black diamond markers indicate performance at no-replay baseline.

was sampled biased by its recency. This meant that rewarded or unrewarded trials could be replayed, and high- or low-RPE trials could be replayed, although there would be a tendency towards higher RPE.

Under this policy, the model performed better (fig. 4.12). With a recency factor of 50, more replays accelerated learning in the initial learning state, the revaluation stage and the reversal learning stage. Effectively this was because sampling from representative trials was equivalent to obtaining more online trials, creating the same effect as more online trials.

RPE-biased state-action-pair replay

Finally, state-action pairs were biased by the average recent reward-prediction error elicited by each pair. A weighted average of RPEs for each pair was calculated, the state-action pairs were chosen probabilistically in proportion to the weighted average, and one recency-biased sample was selected from the exemplar trials of this pair.

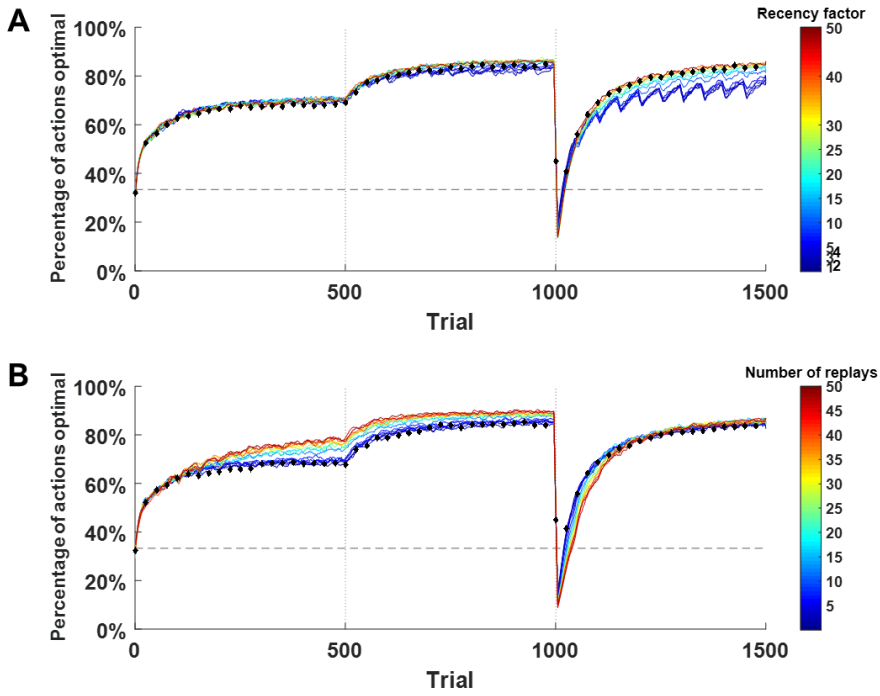

Figure 4.13. RPE-biased state-action-pair replay. Rate of optimal action selection under a policy of RPE-biased state-action-pair replay. **A.** Performance with varying recency factor for skewing replay samples towards more recent trials. **B.** Performance with varying numbers of replay updates per session. Black diamond markers indicate performance at no-replay baseline.

Under this policy, there was substantial improvement not in the speed of learning but in the asymptotic performance reached (fig. 4.13): with a recency factor of 50, the agent converged to a much higher rate of optimal actions under this policy. As with rewarded-state-action-pair replay, the trials sampled from were more representative than previous replay policies, but the preferential learning from surprising information was more efficient at finding weaknesses in the internal representation of the environment and improving them.

4.4. Discussion

A Q-learning algorithm was trained on a simulated version of the stochastic reinforcement learning task presented in Chapters 2 and 3, to test the suitability of the algorithm for modelling rat behaviour and to test the effects of a range of replay policies. The model was able to appropriately solve the task, both in the simple case of an environment with stationary reward probabilities and when presented with non-stationary rewards. In the best case, with stationary rewards and manually tuned parameters, it could perform at close to 100% optimality.

With different parameters, Q-learning showed patterns of learning qualitatively similar to rats' behaviour in Chapter 2: in particular, it replicated the probability-matching tendency to select actions roughly in proportion to their expected reward. This is evidence that Q-learning is an appropriate model of rats' learning on this task, providing a suitable framework for investigating the role of replay in reinforcement learning.

Introducing replay to the model had observable effects on learning, but had the strongest effect when learning parameters were suboptimal. Prioritising single trials for replay based on their recency, reward or reward-prediction error skewed replay towards unrealistic representations of the environment, and impaired learning, similar to what has been reported before (Isele & Cosgun, 2018). Prioritising state-action pairs for replay, and sampling trials from those, was more effective. More specifically, although preferentially replaying rewarded state-action pairs accelerated learning relative to no replay, a stronger effect was seen by replaying state-action pairs with a high average reward-prediction error: the asymptotic performance reached was substantially higher than could be achieved either with no replay or with replay biased towards highly rewarded state-action pairs.

Attempts in machine learning to prioritise surprising or high-reward-prediction-error trials has had mixed results, because, as in these results, it is sensitive to skews which result in an offline training set which is unrepresentative of the environment. Grouping trials of the same type (i.e. state-action pair) to get a weighted average of RPE proved to be a robust way of maintaining environment statistics while identifying weaknesses in the current Q-value representation.

In the neurophysiological literature on replay, RPE-biased replay has not explicitly been reported, although reward-biased replay has (Singer & Frank, 2009). However, reward and reward-prediction error are often conflated in these studies, and evidence that replay in humans can be predicted by neural sensitivity to uncertainty (Momennejad et al., 2018) suggests that reward-prediction error might play a role in biasing replay. This possibility is explored further in Chapter 5.

The results from these simulations suggest that RPE-biased replay might be an effective way of enhancing learning from limited experience which is otherwise suboptimal. To investigate whether this principle applies to reinforcement learning in rats, the hypothesis that RPE-biased replay influences reinforcement learning online was tested by fitting the parameters of a Q-learning model with replay to the behaviour of rats: these results are presented in Chapter 5. The hypothesis that such replay is implemented by a network involving the hippocampus and accumbens is tested in Chapter 6.

Chapter 5: Modelling replay from behaviour

The bulk of this chapter has been published online as a pre-print (Roscow et al., 2019).

5.1 Introduction

Reinforcement learning involves using past experience to guide behavioural policies for future behaviour. In Chapter 2 we saw that rats are capable of achieving this on a probabilistic maze task; in Chapter 3 we saw that learning on this task is associated with activity in the hippocampus and nucleus accumbens, two brain areas known to engage in replay offline; and in Chapter 4 we saw that replay in a simulation of the same task can help an agent achieve better performance, especially when biased by reward-prediction error. Taken together, these results raise the possibility that rats might replay recent experiences on the maze according to reward-prediction error to enhance learning.

Corroborating findings in machine learning replay, activity which is associated with experiences of reward (Foster & Wilson, 2006; Lansink et al., 2009; Singer & Frank, 2009) or fear (Girardeau et al., 2017; Wu et al., 2017), or with recent experiences (Cheng & Frank, 2008), has been found previously to be replayed preferentially in rodents. This suggests a replay bias towards the most salient experiences to be processed, consolidated or incorporated into the internal model of the world. However, these salient experiences could also be interpreted as those with the highest prediction error, i.e. the most informative experiences for updating internal models and for reinforcement learning. Tasks which involve learning the locations of rewards often conflate reward with reward-prediction error (RPE), leading to the possibility that apparent replay biases towards reward actually reflect biases towards RPE.

Here the possibility is explored that it is reward prediction errors, rather than reward or salience, which biases replay. I used variations of a reinforcement learning model, Q-learning, to estimate the value of actions encoded in the striatum during a reinforcement learning task, and varied the amount and type of replay in the model to predict behaviour. In the striatum, representations of reward values differ following learning acquired over weeks compared to when acquired over minutes (Wimmer et al. 2018), and, correspondingly, reward-responsive cells are replayed preferentially in the ventral striatum (Lansink et al. 2009). I therefore propose that replay triggers value updates in the striatum, to enhance striatum-dependent reinforcement learning, and moreover that activity encoding events that resulted in high RPE is preferentially replayed.

Q-learning (Watkins 1989) has been used successfully to model reinforcement learning, particularly in humans (O'Doherty et al., 2003; Daw et al., 2005) but also in rodents (Kim et al., 2013; Ito & Doya, 2009). Q-learning models fit both behavioural outcomes and striatal activity, suggesting that they describe mechanisms of updating values in the striatum in response to RPEs which in turn guide behaviour (Day et al., 2014; Morris et al., 2010; Pagnoni et al., 2002; Roesch et al., 2007). Temporal-difference-based RPEs, i.e. the difference between expected reward and actual reward which drives the update of Q-values, resemble quite closely the dopaminergic input of ventral tegmental area (VTA) to the striatum (McClure et al., 2003; Roesch et al., 2007; Schultz, 2016), which mediates synaptic plasticity in the striatum (Calabresi et al., 2007) and may provide a mechanism of biological implementation of Q-learning. Dyna-Q (Sutton, 2014), a variant of Q-learning which incorporates offline temporal-difference updates, has been used to model replay in ways which produce learning qualitatively similar to animal reinforcement learning (Johnson & Redish, 2005). RPE-biased replay incorporated into machine learning algorithms show that it can also be very efficient, learning to play Atari games (Andrychowicz et al., 2017) or navigate a simulated environment (Karimpanal & Bouffanais, 2017) faster and with more success compared to replay without such a bias.

We trained 6 rats on a stochastic reinforcement learning task which elicited both positive and negative RPE, and fitted Q-learning parameters to each rat's behavioural data. We then included replay events between sessions, to simulate the effect of replay during sleep on reinforcement learning. Four replay policies were compared, prioritising state-action pairs to be updated according to different biases: random replay, replay proportional to expected reward, and two forms of RPE-biased replay. Random replay was included as a control, while reward-biased replay reflects the prevailing view of how replay is prioritised. Fitting the model parameters showed that the two RPE-biased replay policies increased the model's predictive accuracy, while random and reward-biased replay impaired model performance. This suggests that replay between sessions of a probabilistic reinforcement learning task in rats is biased by RPE and not by reward.

5.1.1. Aims of this chapter

1. Train Dyna-Q on behavioural data to infer how replay is biased by reward
2. Generate concrete hypotheses that can be tested with neural data

5.2 Methods

The behavioural task was carried out as described in Chapter 2.

5.2.1. Q-learning

We trained several variations of a Q-learning algorithm on the behavioural data to predict choices of which arm would be entered on each trial. Q-learning is a reinforcement learning algorithm developed for Markov decision processes in which an agent selects actions in its environment and observes the outcome, recording at each time step t its starting state s_t, selected action a_t, resulting reward r_t, and resulting state s_{t+1}. The agent builds up a matrix Q of Q-value estimates for every state-action pair:

$$Q = \begin{bmatrix} Q_{s_1,a_1} & Q_{s_1,a_2} & \cdots & Q_{s_1,a_A} \\ Q_{s_2,a_1} & Q_{s_2,a_2} & & Q_{s_2,a_A} \\ \vdots & & \ddots & \vdots \\ Q_{s_S,a_1} & Q_{s_S,a_2} & \cdots & Q_{s_S,a_A} \end{bmatrix}$$

corresponding to the future discounted expected reward, i.e. the temporal difference between the current state and the reward state. These Q-value estimates are used to guide actions to maximise reward. At each time step t, the Q-value for the state-action pair observed is updated by:

$$Q(s_t, a_t) \leftarrow (1 - \alpha) \cdot Q(s_t, a_t) + \alpha(r_t + \gamma \cdot \max Q(s_{t+1}, a))$$

where $\alpha \in (0,1)$ is a learning rate parameter which determines the degree to which new information overrides old information, and $\gamma \in (0,1)$ is a discount parameter which determines the importance of long-term gains.

In this task, entries into a chosen arm (and arrival at the goal location at the end of the arm) were modelled as actions, while the arm entered on the previous trial, on which reward probabilities were contingent, were modelled as states. Each trial therefore gave rise to one state-action transition out of nine possible state-action pairs.

5.2.1. Q-learning with replay

We used four variants of Q-learning in which additional "offline" updates are performed between "online" trials, based on sequences already experienced, to boost learning. This has the effect of learning from several trials per actual trial of experience, and is similar to the Dyna-Q algorithm which has been shown to speed up learning compared to Q-learning alone (Sutton, 2014) in a manner which may underlie the function of hippocampal replay (Johnson & Redish 2005). Generally, sequences are selected randomly from a memory buffer of recently-acquired experiences, without bias towards any trial or type of trial. Given the observed bias reported in the literature towards salient experiences, such as those rewarded or aversive, we modified Dyna-Q to perform updates only between sessions and to reflect hypothesised biases in four different ways.

5.2.2. Parameter-fitting

Parameter-fitting for Q-learning

First, a Q-learning algorithm (without replay) was trained, to obtain a baseline score against which various replay policies could be compared. Q-values were stored for each state-action pair on the task, and updated according to each animal's experience. A state s_t was defined as the arm visited on the previous trial $t-1$, and an action a_t was defined as the arm chosen on the current trial t. Following each trial of an animal's training, the Q-value $Q(s_t, a_t)$ was updated according to the reward received, $r \in \{0,1\}$ by the Q-learning rule, and Q-values were transformed into a forecast probability of choosing each arm on the subsequent trial.

The learning rate α, discount factor γ, and exploration factor ϵ were free parameters that were tuned to each rat, using the following optimisation procedure. Here we used a reliability score (Murphy & Murphy, 1973), generated based on the forecast probabilities of all trials, to quantify the consistency of the forecast probabilities with the animals' behaviour. The mean observed frequency was calculated for each state-action pair, i.e. the proportion of trials on which a given action was chosen in a given state, and the reliability score R_t for a given trial t was calculated according to:

$$R_t = n_{s_t} \cdot \sum_{a=1}^{n_a} (p_a - o_{s_t,a})^2$$

where s_t is the animal's state on trial t, n_{s_t} is the number of trials on which the animal was in state s_t, n_s is the number of possible actions (3), p_a is the forecast probability for entering arm a, and $o_{s,a}$ is the mean observed frequency of state-action pair s, a. The forecast probability p and observed frequency o were calculated over all trials and sessions, assuming action probabilities that don't change over time. This obscures changes in action probabilities on a faster timescale, as the animals' behavioural policy changed, but employing a smaller time horizon introduces the problem of identifying over what timescale action probabilities are changing.

Parameter optimisation was performed using the reliability error as the cost function. Because the parameter state-space was vulnerable to local minima, and also because it was highly stochastic under replay policies (see description below), a two-step approach was taken to optimise parameters. In the first step, simulated annealing was run 32 times for a maximum of 1000 iterations (or until the reliability error could not be improved by more than 1 x 10^{-6}), using the MATLAB function `simulannealbnd`. Reliability error was averaged over 1,000 runs when computing the cost function, to minimise stochasticity. This function performs a probabilistic variation of gradient descent by taking increasingly smaller steps in random directions, to approximate a global minimum without becoming stuck in local minima. The resulting 32 rough estimates of the optimal parameter values were used as the initial values for the second step: a simple quasi-Newton method using gradient descent, implemented by the MATLAB optimisation function `fmincon`, for a further maximum 1000 iterations. Of the 32 final sets of parameter values, the one which produced the smallest reliability error was used for analysis. All analyses were performed on the average reliability error over 1,000 runs using the given parameter values.

Parameter-fitting for Q-learning with replay

Against the baseline of no-replay, the same optimisation procedure was performed with increasing amounts of replay according to four replay policies. Following each session, a specified number of samples were chosen from all the trials experienced so far. How the samples were selected depended on the replay policy (detailed below); a probability $P(s, a)$ was assigned to each state-action pair to determine which pair to sample from. From the chosen state-action pair, a sample trial was chosen according to the probability $P(i)$ in which a recency parameter ensured that more recent trials were exponentially more likely to be chosen. Q-values were then updated according to the state, action and reward of the sampled trial, in the same manner as "online" Q-value updates described above.

Each replay policy required the same three parameters to be optimised as in Q-learning without replay, plus additional parameters for recency and/or RPE-weighting. Table 5.1 shows the number of free parameters for each replay policy.

Replay policy	Number of parameters
No replay	3
Random replay	4
Reward-biased replay	4
RPE-prioritised replay	5
RPE-proportional replay	5

Table 5.1. Number of Q-learning replay parameters.

These were optimised according to the same procedure as for Q-learning with no replay, described above, for $n = \{1, 3, 5, 10, 15, 20, 30, 40, 50, 75, 100\}$ replay events between each session, resulting in 11 sets of parameter values for each replay policy and each animal. Comparing this to plausible quantities of replay events in animals is not trivial, but studies in which discrete replay events are enumerated report 100-200 bursts of hippocampal activity that can be statistically related to prior experience, over the first one or two hours after experience (Ólafsdóttir et al., 2016; Michon et al., 2019). Separately, reactivation of cell pairs has been found to decay to baseline well within that time period following exposure to familiar environments (Giri et al., 2019), so the first one to two hours is likely to be when most replay of recent experience in a familiar environment occurs.

Random replay

Random replay, biased by nothing but the recency of an action, was included as a control. For each replay event, a state-action pair was chosen at random out of all state-action pairs experienced so far:

$$P(s, a) = \frac{1}{n_{sa}}$$

where $n_{sa}$is the number of state-action pairs experienced (up to 9). The subset of trials experienced, $i \in (1, I)$, which represented this state-action pair were ordered chronologically, and the probability $P(i)$ of a trial i being replayed was determined according to a recency rule with recency parameterφ:

$$P(i) = \frac{i^{\varphi}}{\sum_{i=1}^{I} P_i}$$

Reward-biased replay

Reward-biased replay represents the predominant interpretation of how reward influences replay (Atherton et al., 2015, Carr et al., 2011). For each replay event, a state-action pair s, a was chosen probabilistically in proportion to its Q-value:

$$P(s,a) = \frac{Q(s,a)}{\sum_{s=1}^{n_s}\sum_{a=1}^{n_a} Q(s,a)}$$

The subset of trials experienced which represented the chosen state-action pair were ordered chronologically, and determined according to equation 7.

RPE-prioritised replay

RPE-prioritised replay represents the policy of replaying trials associated with the most surprising outcomes, i.e. where the difference between expectation (Q-values) and experience (reward) was greatest. For each trial t, RPE was calculated as the difference between actual reward and expected reward:

$$\mathrm{rpe}_t = r + \gamma \cdot Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$$

For every trial $i \in (1, I)$ which was an example of a given state-action pair, its absolute value was weighted, determined by a parameter φ raised to the power of its recency i:

$$\mathrm{wrpe}_i = |\mathrm{rpe}_i|\varphi^i$$

The weighted RPEs, wrpe, were then averaged to produce an overall weighted-average RPE, $\mathrm{RPE}_{s,a}$, for each state-action pair s, a, which was more heavily influenced by recent trials:

$$\mathrm{RPE}_{s,a} = \frac{\sum_{i=1}^{I} wrpe_i}{I}$$

The state-action pair with the highest RPE was selected, and the subset of trials experienced which represented the chosen pair were ordered chronologically, and determined according to the recency rule. Once replayed, the rpe_t for the trial sampled was updated to reflect the RPE resulting from the replay event.

RPE-proportional replay

RPE-proportional replay is a variant of RPE-prioritised replay, in which state-action pairs are chosen in proportion to their weighted-average-RPE instead of choosing the pair with the highest weighted-average-RPE. The RPE was calculated as above and a state-action pair to be sampled from was chosen probabilistically according to:

$$p_{s,a} = \frac{\mathrm{RPE}_{s,a}}{\sum \mathrm{RPE}_{s,a}}$$

The subset of trials experienced which represented the chosen state-action pair were ordered chronologically, and determined according to the recency rule. Once replayed, the rpe_t for the trial sampled was updated to reflect the RPE resulting from the replay event.

Shuffling procedure

As an additional control, the parameters were also optimised for shuffled data, in which trial order was randomly permuted 1,000-fold. This preserved the large-scale information in the training data, such as the mean observed frequency and average rewards of state-action pairs and the number of trials in each session between replays, but disrupted the specific structure of how this information was acquired over time.

5.3. Results

5.3.1. Q-learning modelled animal behaviour

We trained a Q-learning algorithm with no replay to generate probabilities of each action for each trial, based on Q-values estimated from the animals' previous experience (fig. 5.1). For each trial, a matrix of Q-values for all state-action pairs was updated based on a rat's experience and used to calculate predicted action probabilities, which were compared to the observed frequencies of state-action pairs to produce a vector of errors for the three available actions. A reliability error was calculated from the summed square of the error vector, weighted by the prevalence of the state. This produced a measure of how reliably the Q-value estimates predicted behaviour (fig. 5.1; see Methods).

Observed action frequency correlated well with predicted action probabilities (fig. 5.2A), indicating a good baseline model for reinforcement learning. Predicted action probabilities from all trials were pooled together and binned in 100 percentile-bins for each animal, and for each bin the mean observed frequency of these actions occurring was compared to the mean predicted probability, resulting in a strong correlation (r =

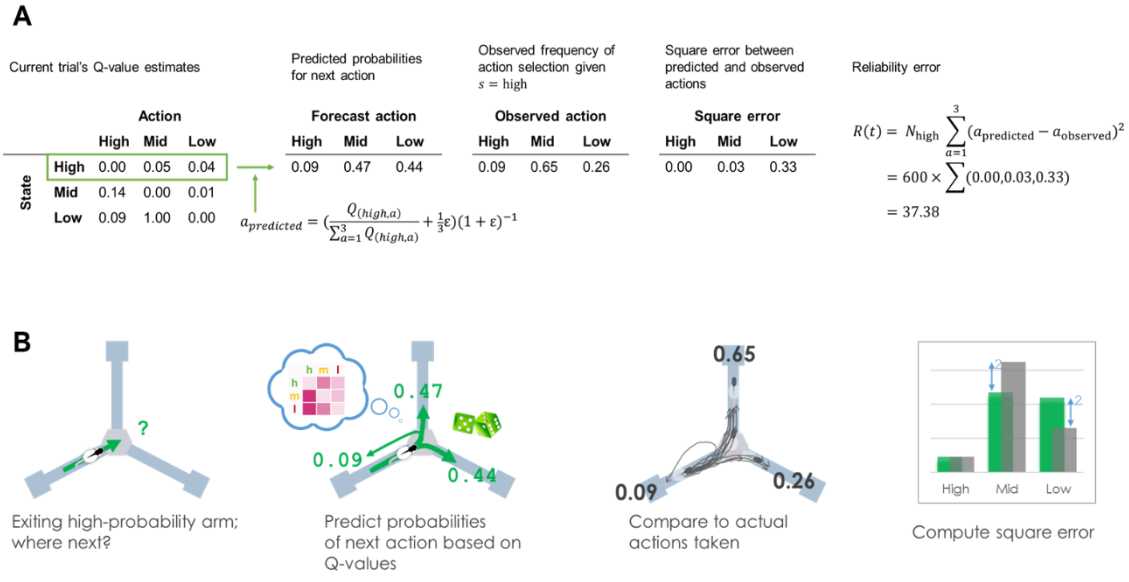

Figure 5.1. Example of model prediction for one trial. Example of model prediction for one trial, t = 100, in which rat H had most recently visited the high-probability arm (s = high) and chose the mid-probability arm (a = mid). **A.** The far left table shows the Q-learning model's estimate of the Q-values based on rat H's experience to date. Other tables show the predicted action probabilities calculated from the Q-values, the ground-truth of observed action frequencies over all visits to this state, and the mean square error between them. Far right shows how the error for this trial is calculated. **B.** A cartoon illustration of the same trial.

	α	γ	ε	Reliability error
Rat H	0.009470	3.340×10^{-09}	0.3451	8.688
Rat I	0.01399	0.2972	0.4035	4.355
Rat J	0.02591	0.5153	0.3173	10.08
Rat K	0.06887	1.000	0.09363	10.66
Rat L	0.6522	1.000	0.3117	18.72
Rat M	0.1345	1.000	0.3137	16.92

Table 5.2. Q-learning parameter values.

0.92, $p = 7.8 \times 10^{-08}$, Pearson's correlation). This result was consistent across animals (correlations ranging from $r = 0.86$ to $r = 0.96$).

The error between predicted action probability and observed action frequency spanned a large range, which was greatest in the earlier training sessions and diminished towards 0 for later training sessions as Q-values were learned (fig. 5.2B; early trials in blue have larger errors).

Reliability errors spanned a different range for each animal (fig. 5.2C), so all further analysis was performed on reliability errors normalised by the mean reliability error for each animal. On this measure, normalised reliability errors were similarly highest in early training sessions, when behaviour is least optimal and most unpredictable. Following this, reliability errors became consistently low for most sessions (fig. 5.2D), confirming a consistent fit with behaviour which captured the learning process over multiple sessions and changes in reward probabilities.

As described in the Methods, the reliability error was used as the cost function to optimise three parameters in the Q-learning algorithm for each animal: a learning rate α, a discount factor γ, and an exploration factor ε. The resulting optimised parameter values are shown in table 5.2. A perturbation analysis was performed to verify that the Q-learning results were sufficiently insensitive to perturbations to the optimised parameter values. At the optimised values, the average normalised reliability error over all trials was, by definition, 1. Perturbing these values by up to 25% in either direction increased the normalised reliability error by less than 0.05 in most cases (fig. 5.2E) and less than 0.1 in all cases, indicating that reliability errors were not overly sensitive to small changes in parameter values.

In summary, the Q-learning algorithm proved able to recapitulate rat behaviour over the course of training and adaptation to new task conditions. The model was robust across a range of parameter values and established a sound basis on which to quantify the effects of mimicking replay by updating Q values between sessions.

5.3.1 Adding RPE-biased replay to the Q-learning model improved prediction accuracy, whereas reward-biased and random replay both reduced accuracy

Against the baseline of no-replay, a variant of the Q-learning algorithm with replay was trained on the same data, with a specified number of samples chosen from all the trials experienced so far to be replayed between each session. Q-learning parameters were optimised for a fixed ($1 \leq n \leq 100$) number of replay events between each session, for each replay policy. All trials experienced by the animal were stored in a memory buffer, and for each replay event a state-action pair was chosen according to the replay policy and a sample trial from this state-action pair was used to update its Q-value. With a random replay policy, all state-action pairs that had been experienced were sampled at random. With a reward-biased replay policy, state-action pairs were sampled in proportion to their Q-values, so that state-action pairs at which rewards had been experienced most frequently would be replayed most. With an RPE-prioritised replay policy, the state-action pair with the highest recent average RPE was sampled. With an RPE-proportional replay policy, state-action pairs were sampled in proportion to their recent average RPE. These latter policies offered two variations on preferentially updating state-action value(s) which had generated the greatest errors, concentrating efforts on correcting the most erroneous expectations of reward.

Compared to the no-replay Q-learning baseline, replay biased by RPE produced a more reliable model of learning, while replay that was random or biased by reward produced a less reliable model (fig. 5.3A; orange and purple compared to blue and green). Both the random and reward-biased replay policies resulted in higher reliability errors ($p = 8.8 \times 10^{-11}$ random, $p = 1.6 \times 10^{-08}$ reward-biased, Wilcoxon signed rank test, Bonferroni-corrected), even with a small amount of replay. Conversely, both the RPE-biased replay polices resulted in lower reliability errors ($p = 6.6 \times 10^{-12}$ RPE-prioritised, $p = 6.3 \times 10^{-10}$ RPE-proportional). This was largely the case for each rat individually (fig. 5.3B). It was true even when one additional sample was replayed between sessions (fig. 5.3D) and remained true when more samples were re-played between sessions (fig. 5.3D-F). Replay of information encoded during trials associated with the most unexpected outcomes therefore significantly improved learning in the model, whereas replay of rewarded trials proved detrimental.

The superiority of the two RPE-biased replay policies was not uniform over the whole training period, however, and two patterns emerged. First, all replay policies showed improvements over no-replay in early sessions, but this effect disappeared in the random and reward-biased policies after roughly the seventh session. This initial superiority of all replay policies over no-replay cannot be due to replay itself because it begins in session 1, before any replay has taken place in the model; rather, it must be due to the non-replay parameters. Specifically, the optimised exploration parameter ϵ was higher in all replay policies than no-replay, so it may be the case that animals tended more towards exploration and relied on Q-values less in early training sessions. The higher ϵ value in the replay policies therefore better modelled behaviour in early

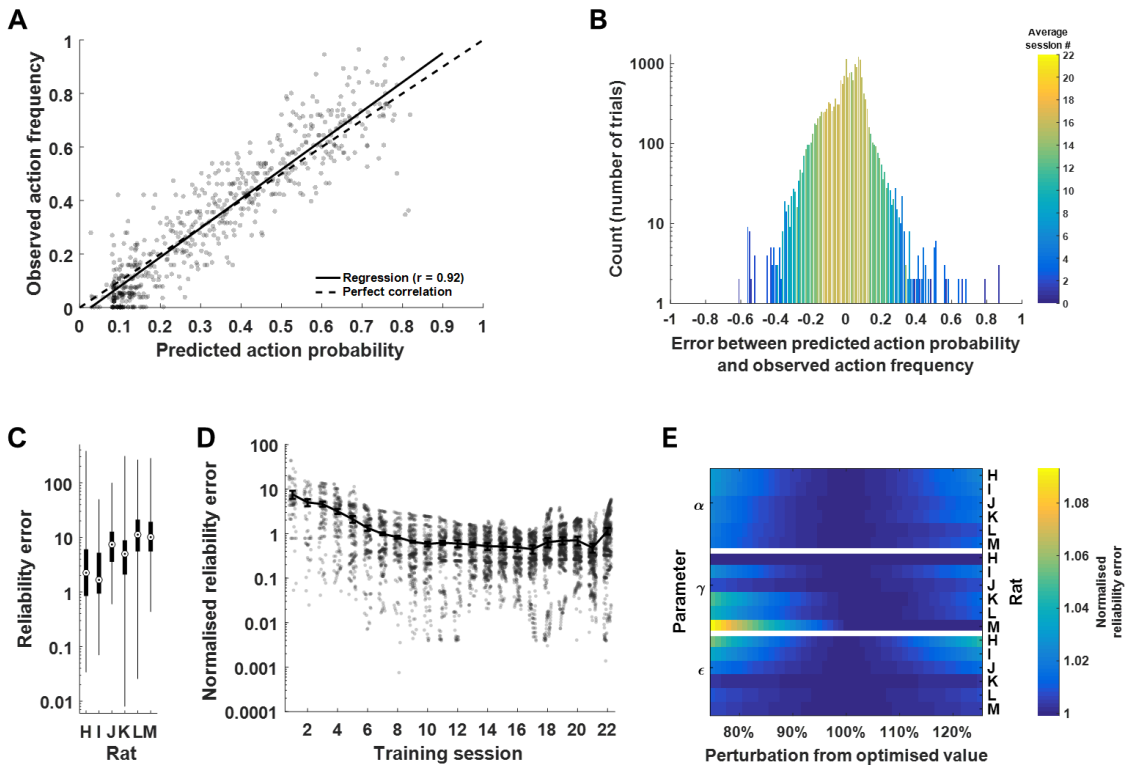

Figure 5.2 Reliability errors. **A.** Reliability diagram (trials pooled across all animals). Observed frequency indicates how often an action was chosen by the animal, averaged over similar predicted action probabilities. Data points indicate mean of each bin. Solid line represents regression (r = 0.92, p = 7.8x10^{-244}); dashed line indicates perfect correlation. **B.** Histogram of residuals of the data in A. Colour scale indicates on average what session the residuals within each bin occurred in. **C.** Range of reliability errors (calculated from residuals) for each animal. A reliability error of 0 reflects perfect modelling of action choices. Boxes represent 25th and 75th percentiles, circles represent median. **D.** Reliability errors for each trial grouped into training sessions, normalised to the average reliability error for each animal (shown in table 1). Data points show normalised reliability error for all trials; solid line represents mean for all animals. Error bars represent s.e.m. **E.** Change in reliability error, normalised to the optimised reliability error for each animal, with varying perturbations to the optimised parameter values. The optimised values for learning rate α, discount factor γ and exploration factor E were individually perturbed by 1%-25% above and below the optimised value and the Q-learning algorithm was trained on behavioural data according to the perturbed parameter values 1,000 times to obtain an average.

sessions, whereas the differences in Q-values resulting from different replay policies impacted behaviour only later.

The second notable pattern is the fluctuations in the reliability errors over training sessions. In the no-replay baseline, reliability error increased in sessions 18-20 and in session 22 (t = 3.54, p = 1.8x10^{-3}, t-test compared to reliability error in sessions 15-17 and session 21). This mirrors an increase in optimal

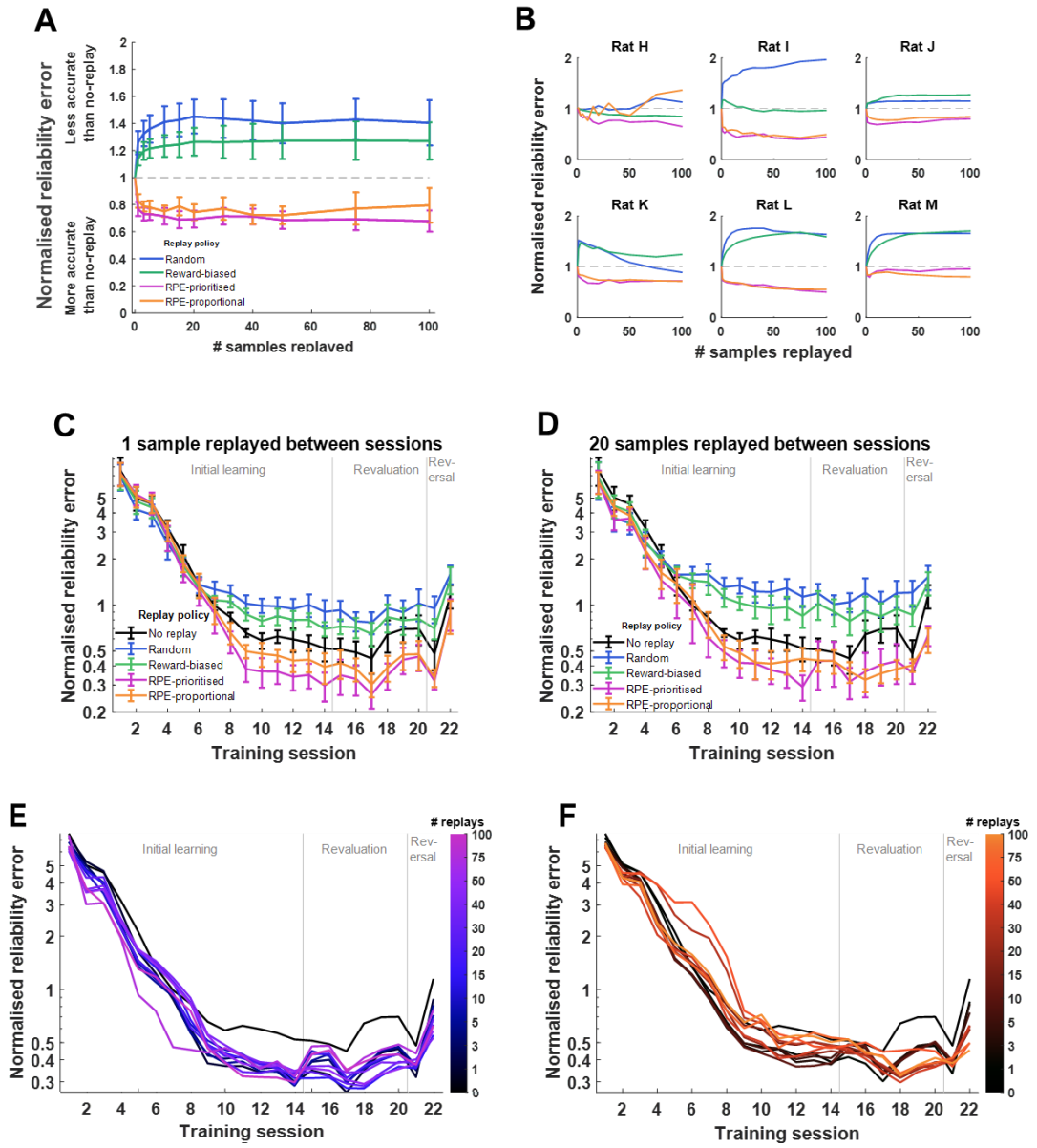

Figure 5.3. Normalised reliability error with replay. A-B. Normalised reliability error with varying numbers of samples replayed between sessions, averaged over all trials, according to the four replay policies shown. Reliability errors normalised to the average reliability with no replay, for each animal. Dashed line represents baseline with no replay. Averaged over all animals (A.) and for each animal individually (B.) C-D. Average reliability error for each session, normalised to the average reliability error for no-replay for each animal. With 1 sample replayed between each session (C.) and 20 samples replayed between each session (D.). Error bars represent s.e.m. E-F. Average normalised reliability error for each session, with varying numbers of samples replayed. E. RPE-prioritised replay policy. F. RPE-proportional replay policy.

behaviour in these sessions during the revaluation stage and reversal stage respectively, suggesting that the model failed to capture subtleties in the learning pattern at these points when animals were adapting their behaviour to changes in reward probabilities. As animals re-evaluated the state-action pairs in sessions 18-20 and adjusted their behaviour accordingly, replay by any policy was sufficient to overcome the increase

in reliability error seen in the baseline, so there was no increase at these sessions (fig. 5.4C; $p = 0.37$ for random replay, $p = 0.94$ for reward-biased replay, $p = 0.081$ for RPE-prioritised replay, $p = 0.06$ for RPE-proportional replay with 20 samples replayed, sessions 18-20 compared to sessions 15-17). This may reflect the faster learning enabled by replaying recently experienced trials. However, as animals reversed their behaviour in session 22, requiring a substantial update to Q-values and a dramatic change in behaviour, increased random replay or reward-biased replay did not improve reliability error. With increased RPE-prioritised or RPE-proportional replay, on the other hand, increasing replay had a particularly strong effect on improving reliability error in session 22 (fig. 5.3E-F). This raises the possibility that RPE-biased replay is especially important for behavioural flexibility of the kind seen in the reversal learning stage.

Rats showed different behavioural phenotypes, as discussed in chapter 2, with regards to their preference for high-, mid-, and low-probability arms (fig. 5.4A). This may relate to irregularities in the rewards received at each arm owing to under-sampling of state-action pairs, which would also extend to the models of behaviour trained on the same actions; differences in the parameters of their learning such as learning rate, which are optimised by the models; and/or inherent preferences for one arm over another, which are not captured by the models. Therefore the predicted arm choices generated by the model under conditions of no-replay (fig. 5.4B), random replay (fig. 5.4C), reward-biased replay (fig. 5.4D), and RPE-biased replay (fig. 5.4E-F), each with 15 samples replayed between sessions, were compared to elucidate some clues about the idiosyncrasies of these learning styles. Predicted action choices were generated in proportion to the predicted action probabilities, run 1,000 times, and averaged over all runs. No apparent systematic differences in predicted behaviour were apparent qualitatively between rats which might explain these behavioural phenotypes, suggesting that learning styles were influenced by inherent preferences that were not captured by the models. Notably, the model of rat M's behaviour with both RPE-prioritised and RPE-proportional replay predicted an aberrantly high preference for the high-probability arm (fig. 5.4E-F) which was not apparent in either the actual behaviour of the rat (fig. 5.4A) or the other versions of the model (fig. 5.4B-D). This accords with the very low improvement of these two replay policies compared to baseline for rat M (fig. 5.3B), and suggests this rat may be an exception to the rule that replay is biased by RPE.

5.3.3. RPE-biased replay did not improve predictions when trained on shuffled data

Given the indication that replay might play different roles in different learning stages, it is important to control for the possibility that parameter values were optimised for the general statistics of rewards and actions in the task, rather than truly modelling the learning curve. Otherwise, the apparent superiority of RPE-biased replay may result from anomalous irregularities in the learning patterns and not true cognitive

Figure 5.4. Predicted action probabilities. A. Actual frequency of arm choices taken by each rat, averaged over trials in one session; this figure is identical to fig. 2.2. B-E. Predicted action probabilities generated by the optimised model of each rat's behaviour according to policies of no replay (B), random replay (C), reward-biased replay (D), RPE-prioritised replay (E), and RPE-proportional replay (F). Each replay policy is based on replaying 15 samples between each session. Green lines indicate entries to the arm that was initialised as high-probability, yellow mid-probability, and red low-probability.

processes. Therefore, the same algorithms were trained on shuffled behavioural data in which the order of trials was randomly permuted 1,000-fold. This preserved the average frequency of state-action pairs and their associated rewards, as well as the lengths of training sessions, but altered the learning curve including revaluation and reversal learning.

Overall, the reliability errors for Q-learning with no replay were lower for shuffled data than real data, because shuffled behaviour was necessarily more consistent over time and therefore more predictable. Similarly to real data, reliability errors decreased sharply in early training sessions before reaching an asymptotic level

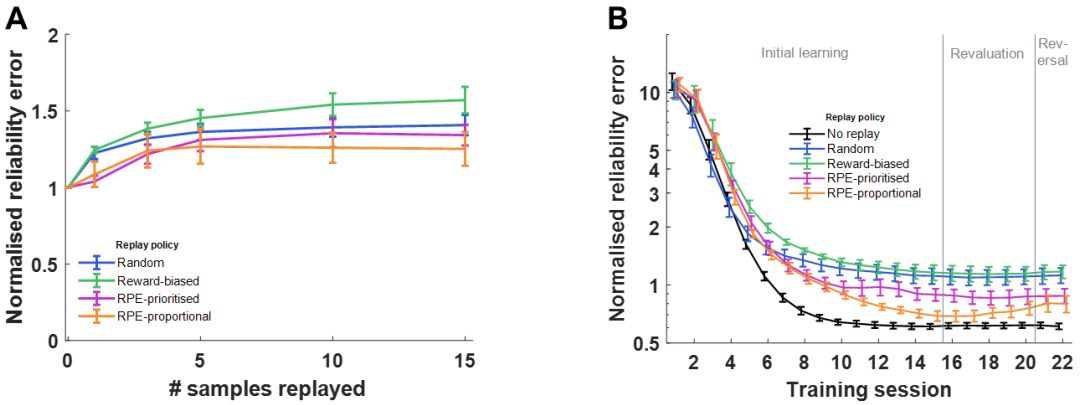

Figure 5.5. Normalised reliability error for shuffled data. A. Normalised reliability error with varying numbers of samples replayed between sessions, trained on shuffled data in which trial data (state, action and reward) are randomly permuted. Dashed line represents baseline with no replay. B. Average reliability error for each session of shuffled data, normalised to the average reliability error for no-replay for each animal, with 15 samples replayed between each session. Error bars represent s.e.m.

(fig. 5.5), because Q-values in early training sessions were distorted by unrepresentative rewards as a result of a small sample size of trials experienced. Unlike real data, the approach to asymptotic reliability error was smooth and monotonic.

Crucially, compared to the no-replay baseline, no replay policy improved reliability error. All replay policies resulted in higher normalised reliability errors than no-replay ($p = 6.9\text{x}10^{-6}$ random, $p = 6.9\text{x}10^{-6}$ reward-biased, $p = 1.6\text{x}10^{-5}$ RPE-prioritised, $p = 3.4\text{x}10^{-5}$). This confirms that the improvement in reliability error in the real data is a result of better predictions of the learning process, and not better convergence to general statistics in the task.

5.3.4. Replay-biased RPE was the best predictor for all state-action pairs

We next accounted for the skew in training data towards the state-action pairs that were chosen most frequently. The transition from the high-probability arm to the mid-probability arm and vice versa (as they were in the initial and revaluation learning stages) were the most commonly experienced state-action pairs, representing 42% of trials overall, and the reliability error was weighted by the frequency of each state such that errors in the more common states contributed more to the overall reliability error than errors in the less common states. We therefore confirmed that Q-learning with RPE-biased replay learned to correctly predict all actions and not just the more-frequently chosen actions to which the cost function was skewed.

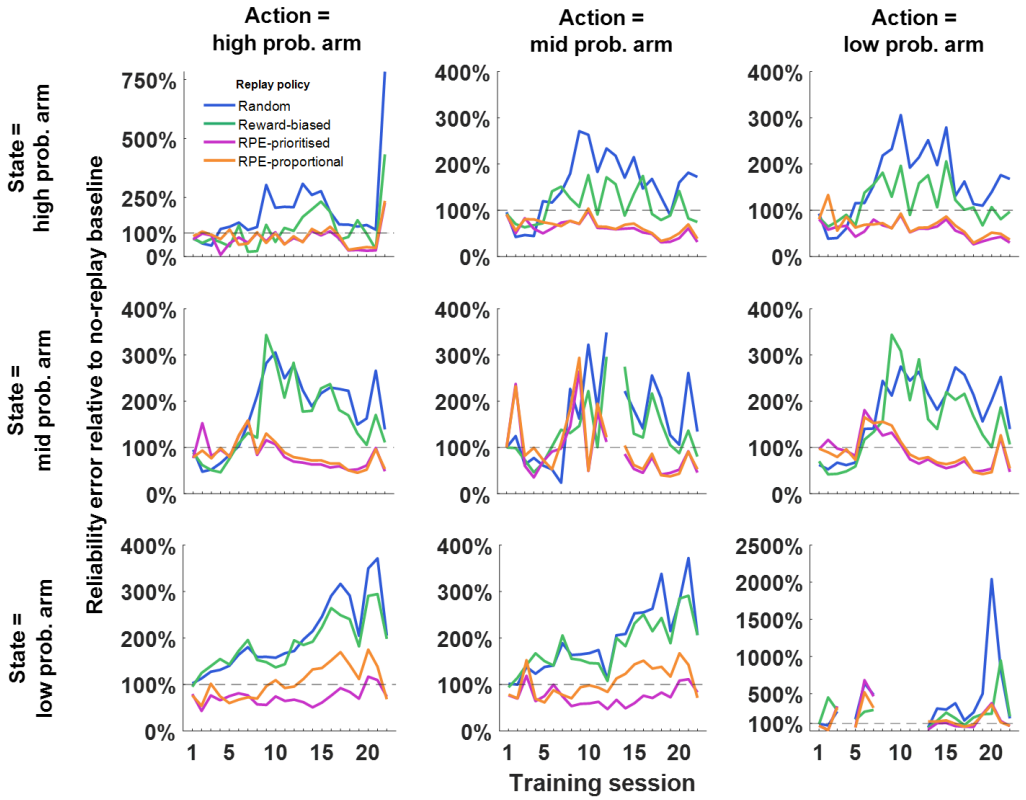

Figure 5.6. Change in reliability error for all state-action pairs. Change in reliability error for all trials on which a given state-action pair was expressed, with 15 samples replayed, relative to no-replay baseline. Intersection of "State = high prob. arm" and "Action = mid prob. arm" indicates a transition from high-probability arm to mid-probability arm.

Figure 5.6 shows the improvement in reliability errors for each replay policy over no-replay baseline, for each state-action pair separately. Despite the skew in training data, the RPE-biased replay policies outperformed random and reward-biased replay policies for every state-action pair, although the improvement was not identical in each case. Nevertheless, the broad conclusion can be reached that RPE-biased replay policies better predicted learning than either no-replay, random replay or reward-biased replay for all state-action pairs.

5.1. Discussion

Rats were trained on a reinforcement learning task designed to dissociate reward outcome (presence or absence of reward) from reward prediction error (RPE; an unexpected reward or absence of reward) on each trial. Variations of a Q-learning reinforcement learning model were trained to predict behaviour on the task, and found that Q-learning with replay prioritised by RPE was the best predictor of learning.

The first main result was that Q-learning can suitably model rats' learning of the stochastic reinforcement learning task, producing low reliability-errors when trained on rats' behaviour and predicting the likelihood of actions on each trial. This is consistent with other studies showing that Q-learning can predict behaviour in a range of tasks in rodents, monkeys and humans (Ito & Doya, 2009). Given this result, we then proposed that adding replay to the Q-learning model between sessions might better reflect learning and therefore better predict behaviour. However, under a policy of replaying state-action pairs randomly, this produced higher reliability errors overall, indicating a worse model of the cognitive processes underlying reinforcement learning. Similarly, biasing replay by sampling from state-action pairs which had produced the largest recent reward also increased reliability errors relative to no-replay.

In contrast, biasing replay by sampling from state-action pairs which had produced the largest recent RPE decreased reliability errors. From this we conclude that the cognitive processes involved in the learning of this task are influenced by offline activity that takes place between sessions. Performance on memory tasks has widely been found to improve following a period of sleep (Stickgold, 2005; Marshall & Born, 2007; Diekelmann & Born, 2010), associated with replay of activity which encodes recent experiences during hippocampal sharp-wave ripples (Ólafsdóttir et al., 2018). We therefore propose that such offline replay underlies the RPE-biased offline updating of state-action values which influenced reinforcement learning in this task.

The suggestion that hippocampal replay might be biased by RPEs differs from the commonly held view that replay is biased by reward itself (Ambrose et al., 2016; Atherton et al., 2015; Gruber et al., 2016; Singer & Frank, 2009). However, the studies on which this conclusion is based generally do not use tasks which explicitly dissociate reward from RPE, so these results in the literature are not inconsistent with our suggestion that RPE biases replay.

Our conclusion that RPE-biased replay (but not random or reward-biased replay) improved model predictions is strengthened by the fact that this result did not hold when training data were shuffled. When the trial order was shuffled, such that there was no correlation between learning and behaviour, all replay policies produced higher reliability errors in predicting the animals' behaviour. This means that the influence of RPE is a feature of the learning process and not an epiphenomenon resulting from the general statistics

of behaviour. Moreover, the result did hold for all state-action pairs, despite the overrepresentation in training data of those most frequently experienced. This gives credence to the notion that the Q-learning model with replay biased by RPE is a good overall model of state-action values held by the brain.

Despite the prevalence of the idea that reward biases replay, our alternative theory that RPE biases replay fits better with existing research on the role of dopamine. Dopaminergic projections from the ventral tegmental area (VTA) to CA1 in the hippocampus have been found to modulate both replay during sleep following exposure to a novel environment, and subsequent memory performance in the same environment (McNamara et al., 2014). It is suggested that dopaminergic neuromodulation might tag synapses by upregulating plasticity-related proteins, causing long-lasting potentiation which allows the stabilisation of the memory trace during subsequent sleep and rest (Frey & Morris, 1998; Redondo & Morris, 2011). Phasic dopaminergic inputs to the hippocampus are triggered not only in response to novelty, but also in the context of reward (Schultz et al., 1997), offering a likely mechanism by which reward-related information might influence replay. Indeed, post-task replay has been found in reward-related VTA cells (Gomperts et al., 2015; Valdés et al., 2015). However, such phasic dopamine activations are typically elicited in response to anticipation of reward and RPEs rather than reward itself (D'Ardenne et al., 2008; Dayan & Niv, 2008; Montague et al. 1996; Schultz 1998; Schultz et al. 1997). These phasic dopamine signals could therefore bias hippocampal replay towards activity associated with RPEs; it is less clear how activity associated with reward per se might bias replay.

Several studies have expressly linked replay to reward, ostensibly in contrast with our results, but often RPE is a confounding factor in these which cannot be discounted. In humans, high monetary reward (but not low monetary reward) is linked to sleep-dependent improvements in associative memory (Igloi et al., 2015; Studte et al., 2017); in this task RPE was not estimated but would presumably be higher overall in the high-reward than low-reward condition, conflating reward-dependent effects with RPE-dependent effects. In rodents, newly-rewarded behaviour has been associated with replay more than behaviour which had been rewarded in previous sessions (Singer & Frank, 2009); here, the authors attributed this replay bias to novelty, but it is also consistent with in-creased RPE when new behaviours are rewarded for the first time. Moreover, following extended reinforcement of both behaviours, the replay bias for the newly-rewarded behaviour was eliminated. In a third study, results were more mixed: following an increase in reward magnitude at one end of a linear track, there was more replay associated with the larger-magnitude end than the unchanged-magnitude end, correlated with both reward and RPE (Ambrose et al., 2016). However, following an elimination of reward at one end, there was a reduction in replay following a reduction in reward despite the increase in RPE. This is more consistent with reward-biased than RPE-biased replay, although the authors noted a rebound effect when the eliminated reward was reinstated: greater replay was found at the reinstated-reward end than the unchanged-reward end, despite identical reward magnitudes. This leaves open the possibility of bias by positive over negative RPEs. A fourth study found more replay of large-reward-related activity than small-reward-related activity on a maze task (Michon et al., 2019), but because

reward was received on every trial analysed, any effects of reward magnitude are conflated with positive reward-prediction error.

Conversely, the specific case for RPE-biased replay is supported by findings that neural sensitivity to RPEs in humans predicts the amount of awake replay during a reinforcement learning task, and replay amount correlated with subsequent performance in a task requiring behavioural flexibility (Momennejad et al., 2018).

In addition to human and rodent studies, findings from the literature on machine learning show some consistency with our results. A number of machine learning studies have found that storing new information in memory buffers and sampling from it at regular intervals, similar to hippocampal replay, can speed up learning (Lin, 1992; Mnih et al., 2013, Mnih et al., 2015), and more so when replay is biased by prediction errors (Cichosz, 1999; Schaul et al., 2016). RPE-biased replay may therefore represent an adaptive focus whereby resources are focused on areas of a cognitive model which needs updating.

This model assumes that a cache of all experience is stored from which to be sampled, which is expensive and unrealistic at large scales. This may not be necessary if memory for individual trials is gradually forgotten and subsumed into cortical long-term memory, for example over the course of hours over which cell assembly activation decays (Giri et al., 2019).

Finally, this model leaves open some questions. It will be necessary to directly test this theory by recording neural data from which replay can be directly observed, comparing replay of reward-associated activity with that of RPE-related activity in the VTA or striatum. There is also an open question about possible diverging roles of replay during behaviour compared to prolonged rest and sleep. Here we have considered replay between sessions, which is likely to take place at least partly during sleep; but replay during wake has also been shown to be necessary for learning (Jadhav et al., 2012).

In summary, we found that a Q-learning-based reinforcement learning model which assumes offline updates between sessions is a better predictor of learning behaviour than one which does not assume offline updates. Specifically, this is true when updates are prioritised according to experiences that have recently elicited high RPEs, and not when they are prioritised according to reward or random recent experiences. This finding offers a reinterpretation of how offline activity during rest and sleep might aid reinforcement learning, in terms of RPE rather than reward. In Chapter 6 the hypothesis that activity in the hippocampus and nucleus accumbens underlies such RPE-biased replay is directly tested.

Chapter 6: Replay in hippocampus and nucleus accumbens

6.1. Introduction

Memory consolidation depends in part on plasticity processes that occur during sleep and in a critical period in the hours following learning. Replay, the coordinated activity of cells associated with a new experience during a subsequent rest period, is reported to take place during a similar time window, potentially creating the ideal physiological conditions for synaptic plasticity. In this Introduction, I review the evidence for the role of replay in memory consolidation, reports of replay particularly in subcortical structures, and theories for how replay might be biased towards certain experiences, especially in the context of reward. The Results contain my own data pertaining to the question of how the hippocampus and accumbens might bias replay in the context of reward.

6.1.1. Hippocampal replay

The link between single-unit hippocampal activity during wake and during subsequent sleep was first reported by Pavlides and Winson (1989), who conducted an exploration session with freely behaving rats to identify pairs of non-overlapping place cells, before restraining rats in the place field of one, but not the other, neuron. During subsequent sleep, the place cells which were reactivated during the restraint part of the session showed a higher firing rate and burst rate than the cells which were not reactivated. Subsequent extensions of this behavioural technique identified that correlations between cells, rather than their raw firing rate, increase during sleep (Wilson & McNaughton, 1994); that these correlations arose after behaviour

and were not present to the same degree in pre-behaviour sleep (Wilson & McNaughton, 1994); and that replay occurs during wake as well as sleep (Kudrimoti et al., 1999). Further work discovered that replay is present not only in the hippocampus but coordinated between hippocampus and other cortical (Qin et al., 1997) and subcortical (Pennartz et al., 2004) brain areas. Interestingly, although many subsequent studies have corroborated Pavlides and Winson's (1989) conclusion that offline activity reflects further "processing of information" acquired during wake, their main finding – that place cells active during experience have higher firing rates during subsequent sleep – has been repeatedly refuted, as hippocampal firing rates remain stable although interactions between cells changes (Wilson & McNaughton, 1994; Kudrimoti et al., 1999).

Replay in animal brains is variously defined as the increased correlation between cells which are coactive during behaviour (Wilson & McNaughton, 1994; Skaggs & McNaughton, 1996; Qin et al., 1997; Kudrimoti et al., 1999; Hirase et al., 2001; Pennartz et al., 2004; Jackson et al., 2006); the reinstatement of assembly activity projected in a low-dimensional state (Peyrache et al., 2009; Lopes-dos-Santos et al., 2013; van de Ven et al., 2016); the Bayesian similarity between multi-unit activity during rest and behaviour (Karlsson & Frank, 2009; Kloosterman et al., 2014; Box et al., 2016; Olafsdottir et al., 2018); or the repetition of precise sequences of spikes (Nadasdy et al., 1999; Louie & Wilson, 2001; Lee & Wilson, 2002; Villette et al., 2015). In humans, non-invasive fMRI-based methods compare the similarity between BOLD signal during behaviour or stimulus presentation and BOLD during rest (Deuker et al., 2013; Staresina et al., 2013). These definitions are principally methodological approaches rather than competing theoretical perspectives: in any case, replay is believed to represent the reinstatement of multi-unit activity that encodes past experiences. Central to the theory of replay is that it has molecular consequences for triggering synaptic plasticity between the reinstated cells, and functional consequences for consolidating, and perhaps additionally processing, the represented information.

6.1.2. Role of replay in memory consolidation

Consolidation of new memories appears to depend on a critical period lasting some hours after initial experience, during which brain injury, sleep deprivation, and cognitive, pharmacological or optogenetic interference can disrupt the retention of newly acquired information (Frankland & Bontempi, 2005). This conspicuously coincides with the time course over which the hippocampus has been reported to replay recent memories (Giri et al., 2019). According to the influential complementary learning systems theory, rapid, short-term encoding of new memories in the hippocampus during experience is gradually reinstated into more durable and generalised cortical representations over this consolidation period (McClelland et al., 1995), which depends on processes associated with sleep (fig. 6.1).

Consolidating new experiences from short-term to long-term storage, as well as other purported functions of replay, necessitates interactions between the hippocampus and other brain areas. Correspondingly, local field potential (LFP) and single-unit events are seen in areas which receive hippocampal projections around the time of ripples, serving as a likely mechanism for (or reflection of) systems consolidation. Single-unit activity modulated by ripple times has been found in entorhinal cortex, prefrontal cortex, anterior cingulate cortex, retrosplenial cortex, auditory cortex, parietal cortex, striatum and VTA, which may allow multiple aspects of an experience encoded by disparate brain areas to be reactivated simultaneously (Rothschild, 2019). Low-frequency spindle oscillations in the thalamo-cortical network, which coordinate spiking activity over a large spatial area, are temporally associated with hippocampal ripples during slow-wave sleep (Siapas & Wilson, 1998) and, correspondingly, replay events (Peyrache et al., 2009). The precise timing between spindles and ripples is variable, however: the latency varies between 15ms and 200ms in different parts of cortex and spindles sometimes precede, rather than follow, ripples (Joo & Frank, 2018), suggesting a dynamic coordination of activity between hippocampus and cortex. The cortical transition to a more excitable "up" state during slow-wave sleep also tends to precede ripples (Battaglia et al., 2004), suggesting that excitatory drive to the hippocampus may prompt ripple activity, offering a mechanism for bidirectional flow of information. This is consistent with the observation that presentation of auditory and olfactory stimuli during sleep can bias memory consolidation by targeting the reactivation of associated memories (Ouidette & Paller, 2013), forming a loop in which cortical activity biases the content of hippocampal ripple activity,

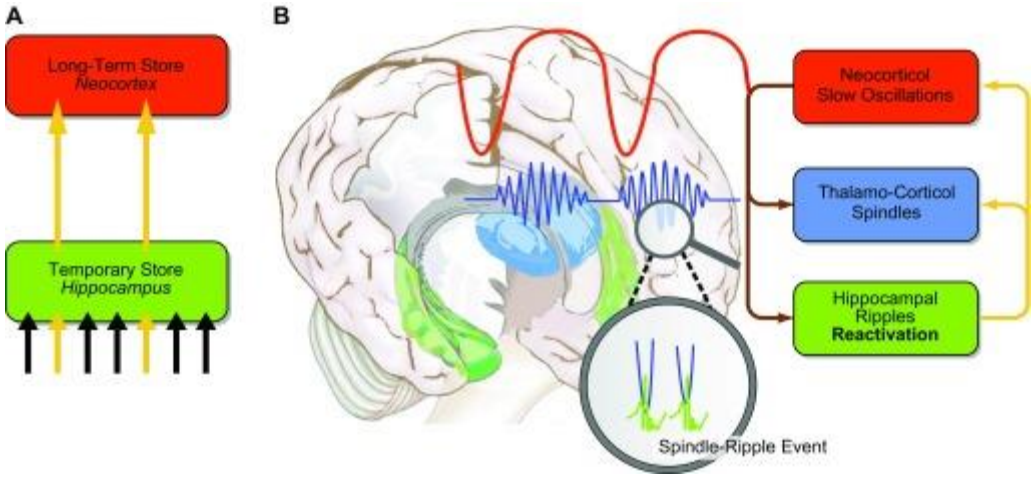

Figure 6.1. Systems-level consolidation during sleep. Neocortical slow oscillations promote thalamo-cortical spindles during the "up" (excitable) phase of the slow wave, and spindles (especially at the transition to an up state) coincide with hippocampal ripples. The temporal entrainment of these oscillations with each other is suggested to allow systems-wide interactions between assemblies of cells that fire during such events. The result is that information encoded in a "temporary store" in the hippocampus is transferred to a long-term store in the neocortex. Although overly simplistic, this framework is consistent with much of the work on sleep-dependent memory consolidation and replay. (Taken from Rasch & Born, 2013.)

which in turn recruits cortical assemblies into the replay event to promote consolidation (Rothschild et al., 2016; Skelin et al., 2019).

The interplay between hippocampus and cortex changes with learning, supporting the idea of information transfer from short-term to long-term memory. Early in learning, the coactivity between prefrontal cortex and hippocampus is correlated between behaviour and sharp-wave ripples (Jadhav et al., 2016); as learning progresses over days, this correlation diminishes (Tang et al., 2017; Joo & Frank, 2018), which is consistent with the observation that cortical representations shift from encoding specific features of experience to more general encoding which is less specific to individual mnemonic episodes (Kitamura et al., 2017; Yu et al., 2017). Representations of new experiences form rapidly in the cortex, but initially require hippocampal input for retrieval; following a consolidation process lasting some weeks, the hippocampus becomes disengaged from retrieval of these memories (Kitamura et al., 2017). However, the view of memory encoding being transferred away from hippocampus and into the cortex is overly simplistic, as some kinds of memory – particularly for individual episodes – continue to require the hippocampus even after extensive consolidation (Joo & Frank, 2018; Yu et al., 2018; Sutherland et al., 2001; Martin et al., 2005). Instead, the organisation of memories may rely on an index stored in the hippocampus and association cortex which unites representations of individual features or elements of a memory stored in disparate areas of the brain, such that when the index is reactivated, various aspects of the memory can be accessed together (Schwindel & McNaughton, 2011).

6.1.3. Replay and sharp-wave ripples

In the short-term, molecular stabilisation of memories in hippocampal circuits involves a cascade of events which promotes protein synthesis to allow structural changes at the synapses, which takes place in a short window of hours after formation of the memory. Consolidating memories across the brain, at a systems level, takes place on a longer timescale and involves processes which rely on synchronous firing across brain areas (Frankland & Bontempi, 2005). This is achieved by projecting bursts of excitatory drive from the hippocampus to other brain areas, sharp-wave ripples, which strongly activate target cells in a wide range of brain areas to allow systems-level interaction.

It is difficult to characterise the precise relationship between replay and ripples, because coincidence with a ripple event is often part of the proscribed definition of replay. Firstly, quantifying discrete ripple events is not straightforward, because it depends on arbitrary parameters and thresholds used for detection which have generally become more liberal over time, producing ripples rates on the order of a few per minute to a few per second (Buzsáki, 2015). Next, quantifying discrete replay events is more complicated still, owing to the stochastic nature of spike timing and the lack of clear definition of what constitutes replay. When

replay events are enumerated they are often identified based on ripples; attempts to decode ripple activity to behaviour have produced estimates from 9% (Michon et al., 2019) to 25% (Wikenheiser & Redish, 2013) of post-task ripples, likely depending on the specifics of data analysis as well as the duration of post-task rest and the demands of and familiarity with the task.

It is important to note that despite the prevalence of correlation-based methods for replay detection, correlations between activity in wake and sleep, i.e. the preservation of a population code, do not necessarily imply a mechanism of learning and memory: existing variations in synaptic connectivity between cells make some patterns of activity more likely than others regardless of any consolidation processes, and there is evidence that new encoding is mapped onto such pre-existing network organisation (Dragoi & Tonegawa, 2014; Liu et al., 2018). Replay analyses generally use pre-task correlations as a control, so this confounds the search for meaningful experience-dependent associations between awake activity and sleep activity. However, manipulations of proposed mechanisms of consolidation and plasticity – most notably ripples – do suggest a causal role for offline activity in spatial learning. Electrically disrupting sharp-wave ripples, which also disrupts the assembly replay, impairs spatial learning (Girardeau et al., 2009; Ego-Stengel & Wilson, 2010; Jadhav et al., 2012; Michon et al., 2019), apparently by preventing this consolidation from occurring. This follows the correlational observation that ripple rates increase following spatial learning (O'Neill et al., 2008; Cheng & Frank, 2008) and correlate with subsequent memory performance (Ramadan et al., 2009). Ripples in these experiments are used as a proxy for replay, and the experimental challenge of disrupting replay per se, without ripples, has not so far been met. Evoking ripples optogenetically from nothing does not appear to produce meaningful replay, but extending the duration of spontaneously-generated ripples does recruit more cells to an ongoing replay event and improve subsequent spatial memory (Fernandez-Ruiz et al., 2019). Similarly, the replay content of ripples can be biased towards some experiences over others by presenting sounds which, having previously been paired with spatial locations, cause a relative increase in replay encoding the associated location (Bendor & Wilson, 2012). Ripples arise in CA3, a hippocampal subregion upstream from CA1 (albeit possibly assisted by excitation in CA2, Oliva et al., 2016, and dentate gyrus, Sasaki et al., 2018), forming when recurrent connections between pyramidal cells facilitate a gradual, exponential build-up of excitatory activity (Buzsáki & Chrobak, 1995; Nakashiba et al., 2009; Csicsvari et al., 2000; Schlingloff et al., 2014; Gulyas & Freund, 2015). Parvalbumin-positive (PV+) basket cells respond to this excitation by firing at ripple frequency and become phase-locked due to reciprocal inhibition which constrains the periods of excitability in which spikes occur (Schlingloff et al., 2014). These interneurons innervate large numbers of nearby pyramidal cells, so their synchronous firing conveys the same rhythm of oscillating excitation and inhibition onto the wider network of pyramidal cells. When a sufficiently large population of CA3 cells participates in the ripple (roughly 10-20%, Csicsvari et al., 2000), it can be transmitted to CA1 via the Schaffer collaterals, whereby broad excitation of CA1 pyramidal cells and ripple-frequency spiking of interneurons causes the ripple to spread by the same mechanisms. Each ripple therefore arises from a different subset of CA3 cells, which recruits a different subset of CA1 cells depending on the synaptic connections between them, which is crucial to the theory that some

assemblies can preferentially take part in ripples. The fast spiking during ripples has been suggested to create the ideal physiological conditions for long-term potentiation (LTP) or spike-time-dependent plasticity (STDP; Girardeau & Zugaro, 2011; Bliss & Collingridge, 1993), and indeed ripples generated in vitro cause changes in potentiation at intra- and extra-hippocampal synapses (Sadowski et al., 2016; Behrens et al., 2005; Norimoto et al., 2018; Lubenov & Siapas, 2008; Colgin et al., 2004) so preferential recruitment of an ensemble of cells to ripples may be selective in promoting plasticity (King et al., 1999; Ormond et al., 2019).

6.1.4. Biasing replay for preferential memory consolidation

The possibility of preferentially recruiting some cells over others into ripples prompts the question of how prior experience might shape ripple activity to direct metabolic resources towards the most "useful" synapses to potentiate. It is thought that some synaptic "tag" is invoked amongst assemblies which are active during behaviour, especially if they are active in the context of novelty or reward, which promotes preferential plasticity and/or participation in ripples later (Redondo & Morris, 2011; Atherton et al., 2015). Neuromodulation may play a role in this: the hippocampus receives dopaminergic input from the ventral tegmental area (VTA) following reward and also from the locus coeruleus (McNamara & Dupret, 2017; Duszkiewicz et al., 2018), which may prompt molecular changes at synapses during behaviour that causes plasticity-related proteins to be synthesised and captured at the excited synapses later on (Martin & Kosik, 2002; Redondo & Morris, 2011). Nevertheless, the theory of synaptic tagging remains imprecise, and it is not clear what neurophysiological, chemical or behavioural factors influence it.

Dopamine has received particular attention for its possible role in biasing memory consolidation and replay, in part because of its broad innervation of many brain areas associated with learning and memory, including hippocampus, striatum and prefrontal cortex. Burst stimulation of dopaminergic VTA terminals in the hippocampus during experience promote both replay of place cells during subsequent rest and spatial memory (McNamara et al., 2014), suggesting that it may act to tag hippocampal synapses by upregulating plasticity-related proteins, promoting potentiation of the synapses and stabilisation of the place cell ensemble (Frey & Morris, 1998; Redondo & Morris, 2011).

In humans, highly rewarded experiences may preferentially benefit from sleep (Igloi et al., 2015; Studte et al., 2017). Evidence from studies which pharmacologically increase or decrease dopamine levels in healthy controls and patients with Parkinson's disease (who naturally have lower dopamine levels) show that dopamine mediates memory performance over a period of minutes and hours. Inconsistent findings have found that memory for highly rewarded items may be boosted by elevated dopamine levels at the point of initial encoding, suggesting dopamine's involvement in tagging memories (Asfestani et al., 2019), or in the subsequent hours, suggesting a role in consolidation (Feld et al., 2014; Grogan et al., 2015). This

discrepancy may arise because dopamine acts to amplify the difference between important and unimportant information, not to globally enhance memory but to selectively promote salient items and accelerate forgetting of less salient experiences (Castillo Diaz et al., 2019; Isotalus, 2019).

Evidence from BOLD activity suggests that involvement of the hippocampus-accumbens-VTA loop during learning (Adcock et al., 2006), and hippocampus-VTA connectivity after learning (Gruber et al., 2016), are associated with the preferential retention of highly rewarded stimuli, which further suggests that offline consolidation involves not just the hippocampus and cortex, but other subcortical structures too. Although the function of replay is often posited as the consolidation of recent experiences from short-term storage in flexible hippocampal networks to long-term storage in more rigid, slow-adapting cortical networks (McClelland et al., 1995), significant reactivation has been found in several subcortical brain regions which suggest a more general reprocessing of experiences.

6.1.5. Replay in subcortical structures

In addition to the accumbens, replay has been identified in two other subcortical structures in coordination with hippocampus: amygdala and ventral tegmental area (VTA).

In the basolateral amygdala (BLA), replay is found associated with cells which are the most strongly coupled to hippocampal activity. BLA cells whose firing during a spatial learning task is correlated with hippocampal activity, and which show ripple-modulation of their firing rate, show increased reactivation with their paired hippocampal cells during post-task sleep (Girardeau et al., 2017). In the VTA, there is conflicting evidence for (1) cells which are responsive to rewarding and/or aversive appetitive stimuli show reactivation during post-task rest (Valdés et al., 2015), and (2) alternatively that reward-responsive VTA cells engage in replay during wake but not sleep (Gomperts et al., 2015). Replay in VTA may be biased towards aversive stimuli (Valdés et al., 2015).

Previously, significant experience-dependent reactivation of cell pairs has been found within the accumbens (Pennartz et al., Lansink et al., 2008) and between accumbens and hippocampus (Lansink et al., 2009) following exposure to probabilistic maze tasks, as well as reactivation of hippocampal-accumbens assemblies following a conditioned place-preference task (Sjulson et al., 2018). This accumbens replay is generally found to be associated with ripples during quiet rest and slow-wave sleep (Lansink, 2008; Sjulson, 2018; but see Lansink et al., 2009). Activity which increased around reward (Lansink et al, 2008; Lansink et al., 2009; Sjulson et al., 2018), which carried spatial information (Sjulson et al., 2018), and which was closely associated with hippocampal activity (Lansink et al., 2009; Sjulson et al., 2018) was found to be preferentially replayed in the accumbens, suggesting a bias towards replay of the most task-relevant

information. During hippocampal replay events which reflect either reward or non-reward locations, both VTA cells (Gomperts et al., 2015) and accumbens cells (Sjulson et al., 2018) which are reward-responsive during the task preferentially take part in those replay events which reflect reward locations.

From this we can hypothesise a role for accumbens replay in reinforcement learning – consolidating associations between locations and reward in order to guide behaviour – but an explicit link between reinforcement learning and accumbens replay has not been found. In this chapter, the data presented in Chapter 3 were reanalysed, with a focus on activity during post-task rest, to investigate ripple-associated replay of activity within and between accumbens and hippocampus. The results of Chapter 5 predict sleep-dependent memory consolidation biased towards experiences with the highest reward-prediction error; here, the hypothesis is tested that such memory consolidation is underpinned by hippocampal-accumbens replay in the same task.

6.1.6 Aims of this chapter

1. To corroborate previous reports in the literature of replay in hippocampus and accumbens following training on a probabilistic maze task
2. To identify the behavioural correlates during the task described in Chapters 2 and 3 of activity that is replayed during post-task rest
3. To test the prediction made in Chapter 5 that replay in the accumbens is biased by reward-prediction error and not reward

6.2. Methods

Binless spike trains. Spike trains were convolved with a Gaussian kernel of 0.05/√12, approximately equivalent to discrete binning of spike trains into 50ms bins, using a method which avoids the variability associated with discrete binning and improves analysis of spike train correlations and explained variance (Kruskal et al., 2007).

Explained variance. Explained variance (EV) is defined as the square of the partial correlation coefficient which represents the proportion of variation in correlations during POST which can be explained by variation in correlations during TASK beyond what can be explained by variation in correlations during PRE; this is contrasted with the reverse explained variance (REV) based on correlations during PRE as a control. Following Kudrimoti et al. (1999), Pearson's correlation coefficients were calculated between binless spike trains equivalent to 50ms bins (see Chapter 3 Methods), for the PRE, TASK and POST periods separately, and combined to create three correlation matrices. The similarity between PRE, TASK and POST was calculated by taking the correlation coefficient r between their correlation matrices:

$$EV = \left(\frac{r_{TASK,POST} - r_{TASK,PRE} r_{POST,PRE}}{\sqrt{(1 - r_{TASK,PRE}^2)(1 - r_{POST,PRE}^2)}} \right)^2$$

This gives a value bound by 0 and 1. REV was calculated by reversing PRE and POST to obtain a baseline value against which EV could be compared. The significance of explained variance for a given session was calculated by performing a permutation test in which the firing rates for cells were shuffled within the same time bin, which controls for spurious correlations caused by population increase in firing rates. For explained variance calculated between CA1 and accumbens, firing rates were shuffled between cells within the same brain region. This permutation was performed 1,000-fold to form a null distribution of EV-REV scores. Observed EV-REV was considered significant if it exceeded the 99.9th percentile of the null distribution.

Experience-dependent increases in cell-pair coactivity. To ascertain which cell pairs showed a significant increase in correlation between PRE and POST, Pearson's correlation coefficients were obtained for binless 50ms spike trains and concatenated for periods where the rat was immobile for at least 10 seconds. For cell pairs which had a significant correlation during POST ($p < 0.05$, corrected for multiple comparisons with critical value of 0.2), correlation coefficients for PRE and POST were transformed to z-scores using Fisher's method, to obtain a z-statistic for the difference in correlations. Null distributions were obtained by randomly permuting the binless firing rates for one of the pair 1,000-fold and applying the same analysis to produce a distribution of z-statistics. Cell pairs whose z-statistic exceeded both 1.64 (equivalent to $p < 0.05$) and the 99th percentile of the null distribution of z-statistics were designated significant.

To determine the behavioural correlates of these cell pairs, their activity was transformed to a binless 50ms spike train and z-scored, and z-scores below 0 were raised to 0. These transformed spike trains for a pair of cells were multiplied together to give their coactivity, with a lower bound of 0. Coactivity was itself z-scored over the recording session for a given cell pair, to reveal times of high coactivity between pairs of cells.

6.3. Results

The single-unit data presented in Chapter 3, recorded simultaneously from dorsal CA1 and nucleus accumbens in one rat, were analysed again here. Recordings were made over 17 sessions during the learning of a probabilistic maze task, as well as during two hours of rest before and after. Here, the activity of single cells and cell pairs was compared between rest and behaviour to identify what, if anything, was replayed.

6.3.1. Significant explained variance during post-task rest

A number of previous studies have found significant reactivation of correlated activity in spatial tasks during post-task rest, both within the accumbens and between hippocampus and accumbens. First, to confirm whether there was significant replay during post-task rest in these results, correlations between cell-pairs were assessed during the TASK, PRE-task rest and POST-task rest to calculate the degree of explained variance in POST correlations that could be explained by TASK correlations, controlling for PRE correlations (see Methods). Pooling all of sessions 1-17 together, there was an overall average explained variance (EV) of 39.9% and reverse explained variance (REV) of 17.9% for the first 15 minutes of POST compared to the last 15 minutes of PRE for pairs of CA1-CA1 cells (paired t-test, $p = 0.011$), an EV of 34.4% and REV of

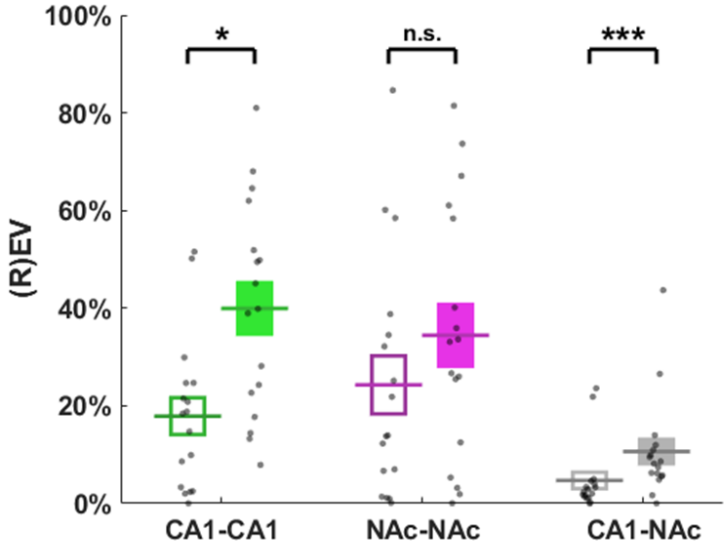

Figure 6.2. Overall explained variance. Explained variance (EV; filled boxes) and reverse explained variance (REV; unfilled boxes) for each session. Intra-area (CA1-CA1 and NAc-NAc) and inter-area (CA1-NAc) cell pairs shown separately. * indicates significance of t-test at $p < 0.05$, *** indicates significance of t-test at $p < 0.001$.

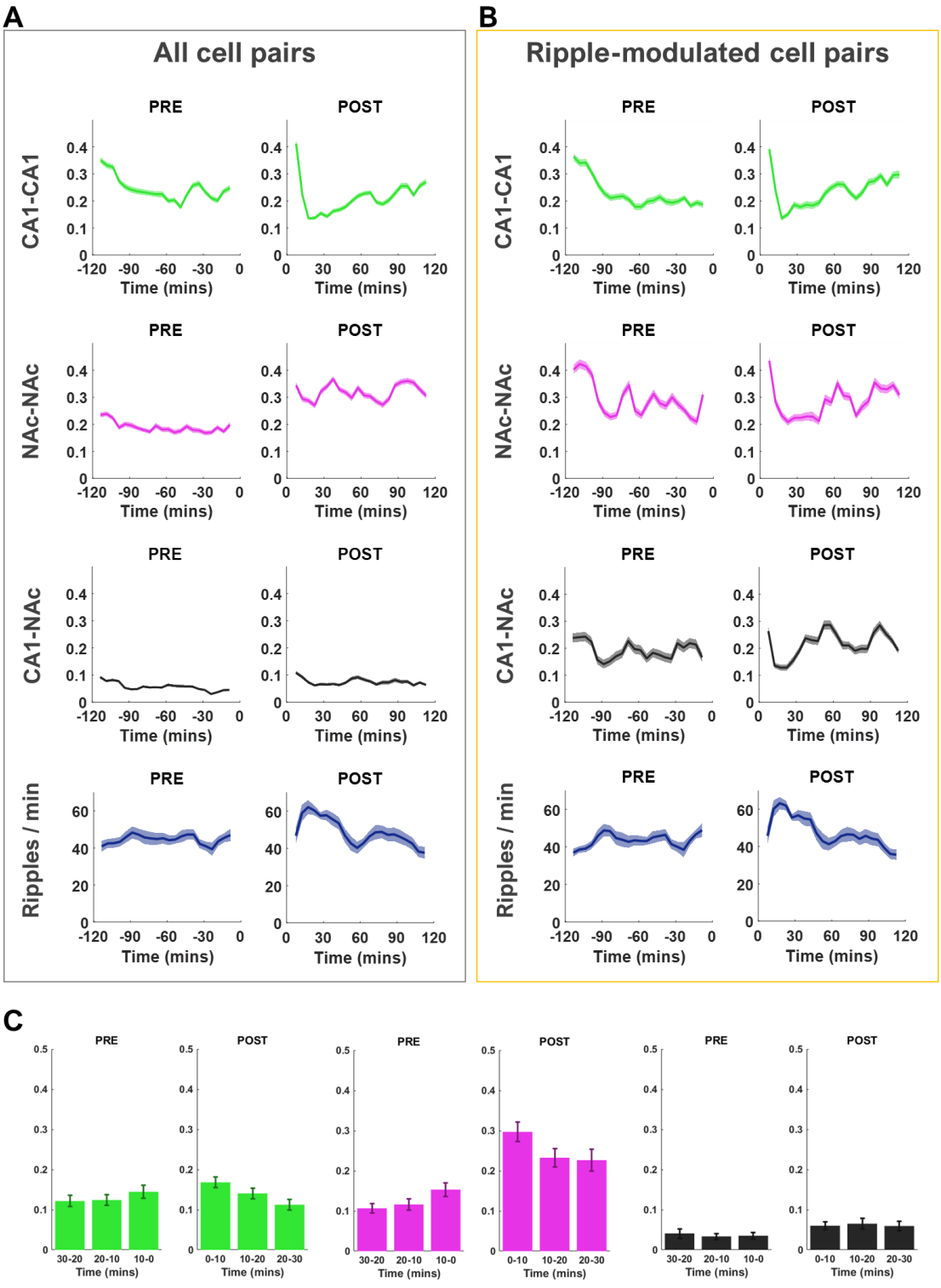

Figure 6.3. Explained variance over time. **A-B.** EV (POST) and REV (PRE) calculated in sliding 15-minute windows over the whole rest period. Two sessions (2 and 16) were excluded because POST periods were less than two hours. **B.** EV and REV calculated from pairs of cells which both showed positive modulation by POST ripples. Bottom: the number of ripples in sliding 15-minute windows. **C.** EV (POST) and REV (PRE) calculated from concatenated periods where rat was immobile for at least 10 seconds.

24.3% for accumbens-accumbens cell pairs, and an EV of 10.6% and REV of 4.7% for CA1-accumben cells pairs (fig. 6.2). Paired t-tests showed that EV was significantly larger than REV for CA1-CA1 pairs and CA1-accumbens pairs, but not accumbens-accumbens pairs (p = 0.011, CA1-CA1; p = 0.34, NAc-NAc; p = 0.00014, CA1-NAc). The significantly larger EV than REV values indicate TASK-dependent patterns of coactivity during POST, i.e. replay.

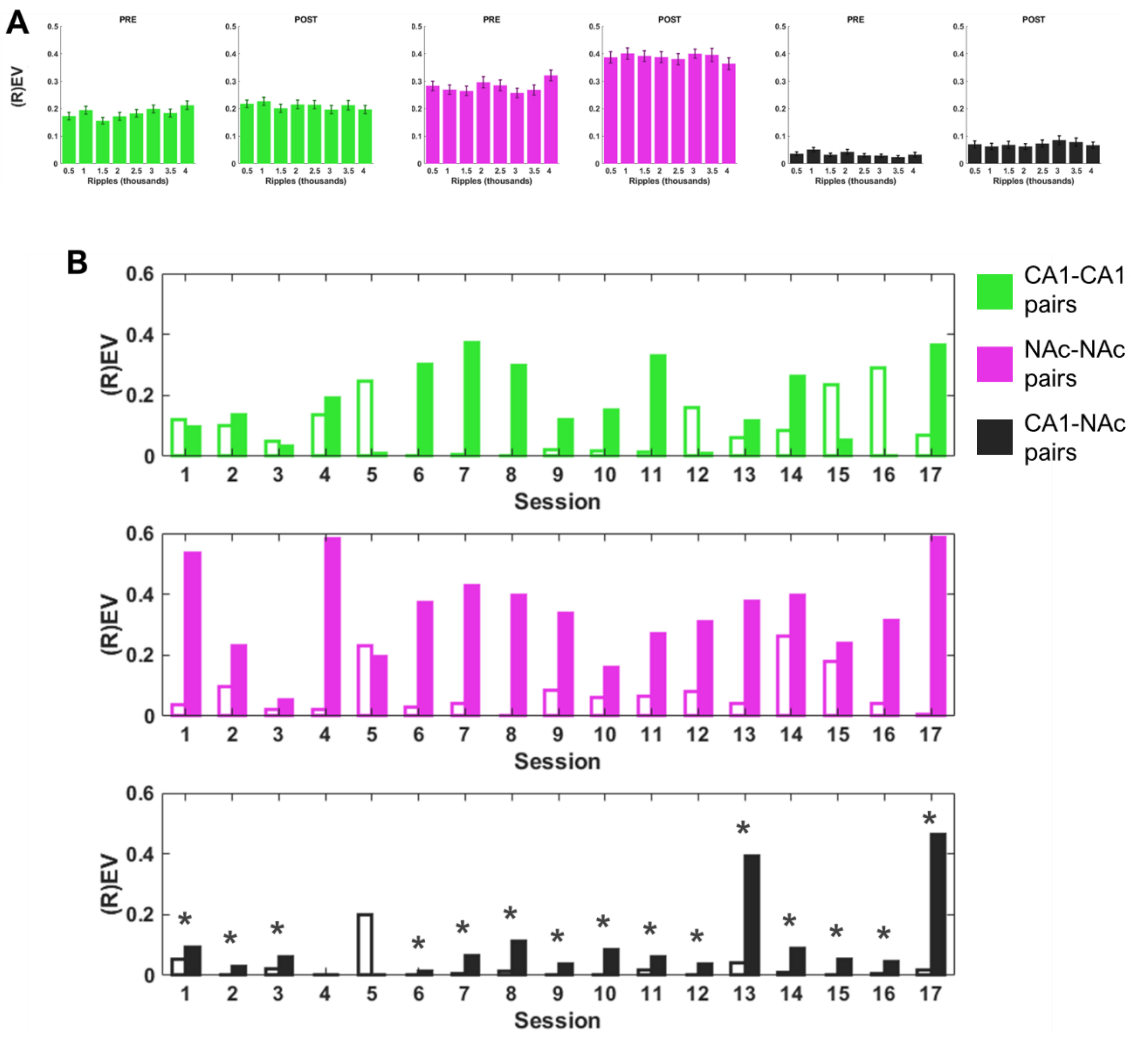

Figure 6.4. Explained variance over sessions. Explained variance (EV; filled boxes) and reverse explained variance (REV; unfilled boxes) for each session, calculated for spiking activity which occurred around ripple times. Intra-area (CA1-CA1 and NAc-NAc) and inter-area (CA1-NAc) cell pairs shown separately. **A.** EV and REV for blocks of 500 ripples. **B.** EV and REV for all ripples in the PRE and POST periods. * indicates significance at p < 0.001 (uncorrected), calculated using permutation test.

Post-task replay in the hippocampus has been found to decay to baseline over a period of roughly 10-30 minutes in well-trained animals or in familiar environments, although no such decay has been reported in accumbens cell-pairs after 30 minutes (Pennartz et al., 2004; Lansink et al., 2008) and hippocampal-accumbens replay has been reported to still be significantly above baseline after 40 minutes (Lansink et al., 2009). Calculating EV and REV in overlapping sliding 15-minute windows, stepping by 5 minutes, showed a similar result here: reactivation of CA1 activity showed a sharp decline after the first 15 minutes, while reactivation of accumbens activity was persistently higher throughout the 2-hour POST epoch (fig. 6.3A). Notably, this was despite the rate of sharp-wave ripples being relatively low in the first 15 minutes of POST (fig. 6.3A, bottom) and when the rat was most reliably awake. Restricting the EV-REV analysis to periods when the rat was immobile for at least 10 seconds – the behavioural state with which both ripples and replay events are most strongly associated – revealed a similar pattern over the course of rest-time, rather than absolute time (fig. 6.3B).

Because the rate of ripples varied over the POST period, EV and REV were also examined over successive ripple times. All sessions contained at least 4,000 ripples during PRE and POST, so EV and REV were calculated within successive blocks of 500 ripples, using the cell-pair activity within 200ms of ripple onset (see Methods; Lansink et al., 2008; Lansink et al., 2009). Accumbens cell-pairs and CA1-accumbens cell-pairs showed persistently higher EV than REV with no evidence of decay (fig. 6.4A). All subsequent EV-REV analysis was based on the concatenated ripple activity. Session-by-session analysis showed significant EV-REV of CA1-accumbens cell pairs on 15 out of 17 sessions (permutation test, 1000 shuffles, $p < 0.001$; fig. 6.4B), compared to 3 sessions for CA1-CA1 pairs and 4 sessions for accumbens-accumbens pairs. All sessions on which there was significant accumbens-accumbens reactivation also showed significant CA1-accumbens reactivation.

Having established that there was significant reactivation within and between CA1 and accumbens, individual cells and cell-pairs which showed signs of reactivation were then analysed for their activity during TASK, to assess the content of what was replayed.

EV-REV analysis was run based on all cell-pairs during ripples (as above) and compared to EV-REV values without each individual cell in turn. A cell which was substantially reactivated would cause EV-REV to decrease when it was excluded from analysis, so the difference between EV-REV with and without each cell was taken as its contribution to EV-REV (Girardeau et al., 2017). Surprisingly, there were no significant associations between ripple modulation, peak firing rate or theta modulation and contribution EV-REV, in either hippocampus or accumbens (fig. 6.5; the same was true for contributions to EV only; no evidence of a bimodal distribution).

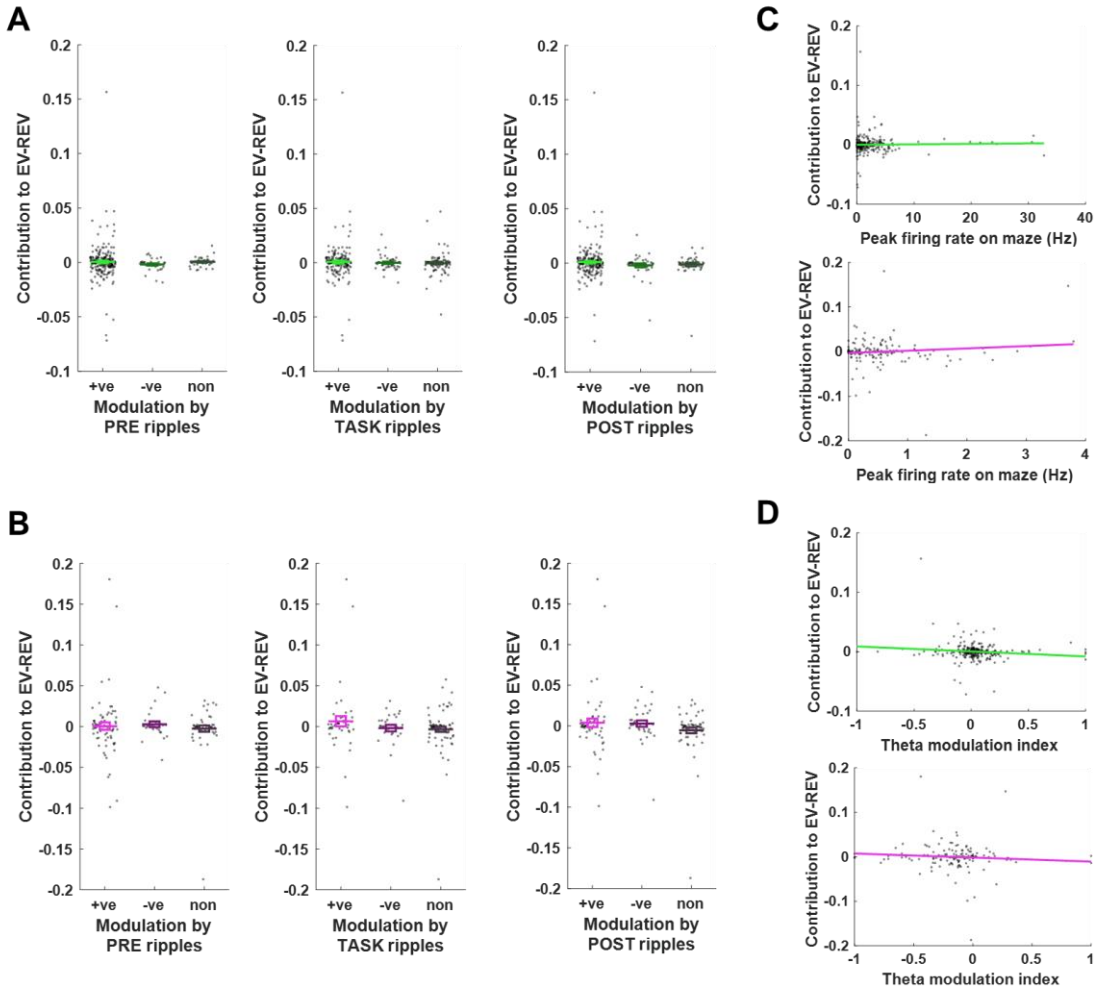

Figure 6.5. Individual cells' contributions to explained variance. A-B. Contributions to inter-area EV-REV per cell, divided by modulation by ripples in each period (PRE, TASK and POST), for CA1 cells (A) and NAc cells (B). C. Correlation between peak per-spatial-bin firing rate during TASK and contribution to EV-REV. D. Correlation between theta modulation index during TASK and contribution to EV-REV.

6.3.2. Behavioural correlates of reactivated cell pairs

To assess which cell pairs exhibited reactivation, cell pairs with significant correlations, and whose correlation increased in POST compared to PRE, were examined.

Binless spike trains for every pair of one CA1 cell and one accumbens cell were analysed for their correlation coefficients during TASK, and compared to a null distribution of correlation coefficients calculated from shuffled spike times (fig. 6.6A-B). Overall, 713 out of 2699 (26.4%) cell pairs showed correlations significantly greater than controls ($p < 0.001$; fig. 6.6C), further confirming a considerable interaction between CA1 and accumbens. Cells which were correlated with at least one other cell did not show greater

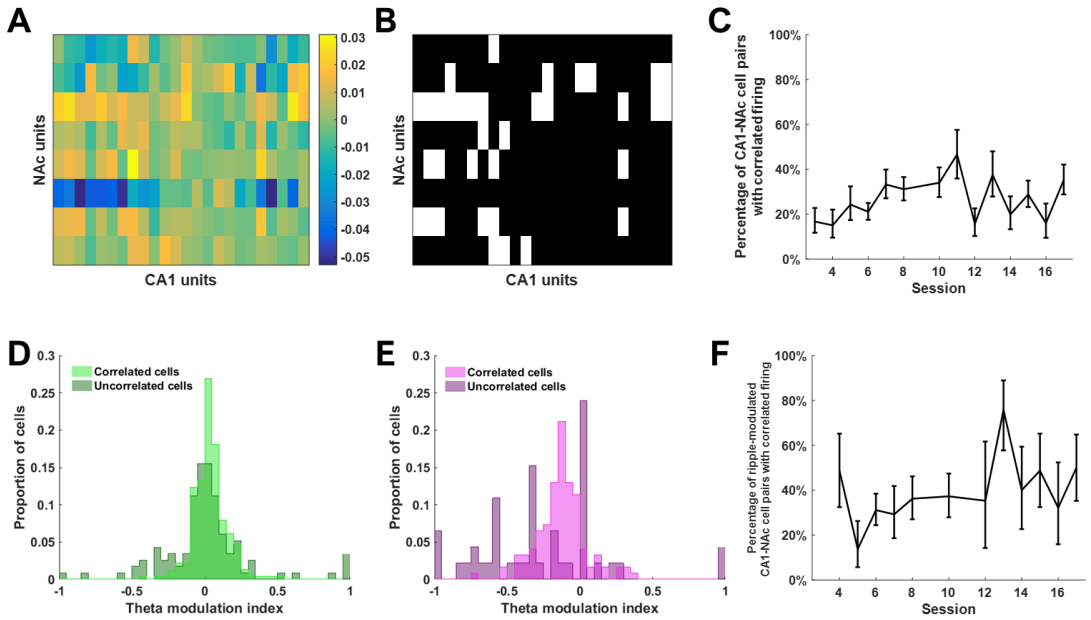

Figure 6.6. Firing rate correlations between hippocampus and accumbens. **A.** Pearson's correlation coefficient between 50ms-binned spike trains during trials, for one example session. **B.** Statistical significance of the correlations shown in A, thresholded after comparison with a null distribution of correlations of shuffled spike trains; significant cell pairs are in white. **C.** Percentage of CA1-NAc cell pairs which showed significant correlations during the task. **D-E.** Theta modulation indices of CA1 (D) and NAc (E) cells which were significantly correlated with at least one NAc (D) or CA1 (E) cell, and cells which were not correlated. **F.** Percentage of CA1-NAc pairs which showed significant correlations, out of a total of cell pairs which were modulated by ripples during POST.

theta modulation than uncorrelated cells (fig. 6.6D-E). Moreover, when the analysis was restricted to pairs of cells that were modulated by ripples during POST, this rose to 32.8% of pairs which were significantly correlated (fig. 6.6F). Strikingly, 100% of NAc cells with significant ripple modulation during POST showed significant correlation with at least one CA1 cell during TASK, further suggesting some involvement in hippocampal replay by accumbens task-related activity.

Next, CA1-accumbens cell pairs were examined for their correlations during PRE and POST. A minority of cell pairs exhibited not only a significant correlation, but a significant increase in their correlation from PRE to POST, indicating systems-level consolidation. 719 out of 8761 (8.2%) CA1-CA1 cell pairs, 183 out of 2427 (7.5%) accumbens-accumbens cell pairs, and 334 out of 4184 (8.0%) CA1-accumbens cell pairs showed a significant increase in their correlation (see Methods). A minority of accumbens cells which showed at least one significant increase in correlation with a CA1 cell were positively modulated by ripples during TASK (48 out of 128, 38%), but this was greater than the proportion accumbens cells with no increase in CA1 correlations which were positively ripple-modulated (15 out of 76, 20%; z-test, $p = 0.0039$).

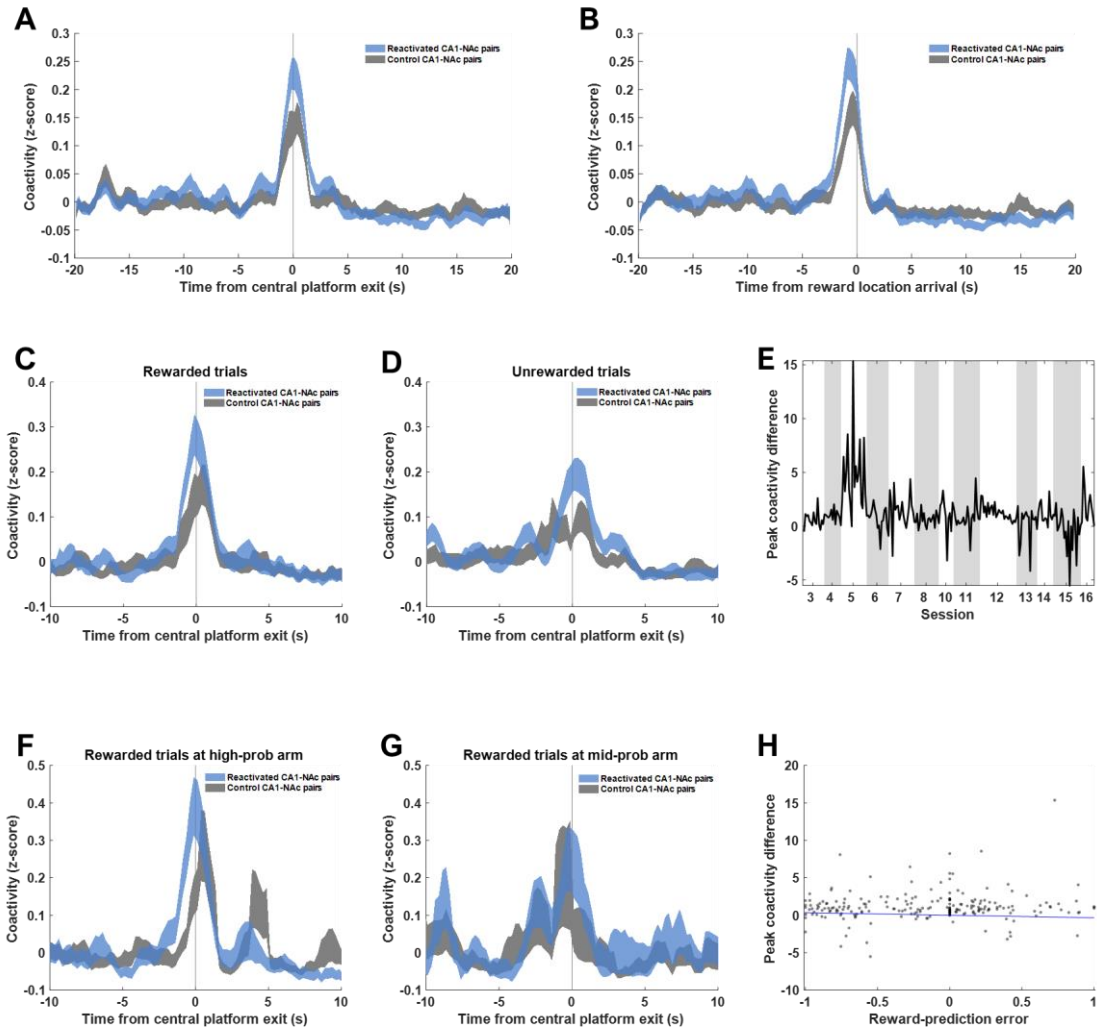

Figure 6.7. Cell-pair coactivity. A-B. Coactivity on all trials analysed between pairs of CA1-NAc cells which showed significant increases in correlation from PRE to POST, as well as control pairs. Trial-averaged and aligned to exit from the central platform (A) and arrival at the reward location (B). C-D. Coactivity on trials which were rewarded (C) and unrewarded (D). E. Difference between coactivity on reactivated pairs and control pairs, at the time point of highest coactivity for reactivated pairs, for every trial. F-G. Coactivity on rewarded trials in a subset of sessions, on the high- (E) and mid-probability (F) arms, reflecting low and high reward-prediction error, respectively. H. Difference in peak coactivity against reward-prediction error for every trial.

These cell pairs which increase their correlation might reflect replay of aspects of task, so coactivity of the significant CA1-accumbens pairs during the task was analysed. Trial-averaged coactivity during the task showed a peak prior to the time of arrival at the reward location, approximately at the point of movement initiation towards the reward location (fig. 6.7A). Population firing rate increases in both areas were apparent at these timepoints (see Chapter 3), and firing rate increases can cause spurious increase in coactivity; to control for this, for every significant cell pair, the coactivity between one of the cells and another non-

significant partner was also calculated. These non-significant cell pairs also showed coactivity around the time of arrival at the reward location, but to a lesser degree and with a later peak (fig. 6.7A-B), indicating that cell pairs whose correlation increased were more active during the trajectory. This increase in coactivity is coincident with the increase in theta-band coherence between CA1 and accumbens (Chapter 3), suggesting theta as a possible mechanism for communication during behaviour which is later replayed.

If there is a replay bias towards experience with high RPE, the activity encoding trajectories which led to a surprising outcome should be replayed more than those which did not. The coactivity analysis was re-run, firstly separating rewarded from unrewarded trials (to assess the effect of reward on replay), and then separating more-surprising rewards from less-surprising rewards. Coactivity was weaker overall, for both reactivated pairs and control pairs, on unrewarded trajectories (fig. 6.7C-D); but reactivated pairs showed stronger coactivity than control pairs on both trial types, so reward outcome did not account for the difference.

Next, sessions on which the rat's performance was good were examined to assess the impact of reward-prediction error on coactivity. (Specifically, sessions 1-4 were excluded because task rules were still being acquired; sessions 13-17 were excluded because reward probabilities changed; and session 9 was excluded because tracking data were not available, leaving 7 sessions for this analysis.) Rewarded trials on the high-probability arm presumably elicited a lower RPE than rewarded trials on the mid-probability arm in these sessions, so the coactivity of significantly reactivated cell pairs on these trial types was compared. Contrary to expectations, coactivity was higher on rewarded trials at the high-probability arm (fig. 6.7F).

To more precisely analyse the relationship between reward-prediction error and coactivity, the difference in peak coactivity on the trial for significant and control cell pairs was obtained for each trial. Reward-prediction error was estimated using the Q-learning method described in Chapter 5. Although there was some evidence that coactivity varied over learning (fig. 6.7E), there was no correlation between coactivity and RPE (fig. 6.7H; Pearson's correlation coefficient, $p > 0.05$).

CA1 and accumbens cells are known to have different firing properties modulated by place, running speed, and proximity to reward. So to further characterise the encoding of replayed activity, the firing of cells which exhibited significant increases in correlation with at least one other cell was compared to the firing of cells which did not exhibit any such replay. In CA1, reactivated cells showed a higher firing rate overall during trials than their non-reactivated counterparts, with a sharp peak prior to exit from the central platform (fig. 6.8A). In contrast, their reactivated accumbens partners showed a peak in firing approximately 1.5 seconds later, ramping up their firing towards the point of arrival at the reward location before sharply dropping off (fig. 6.8B). A similar pattern was seen in raw (not z-scored) firing rates (fig. 6.8C-D), although

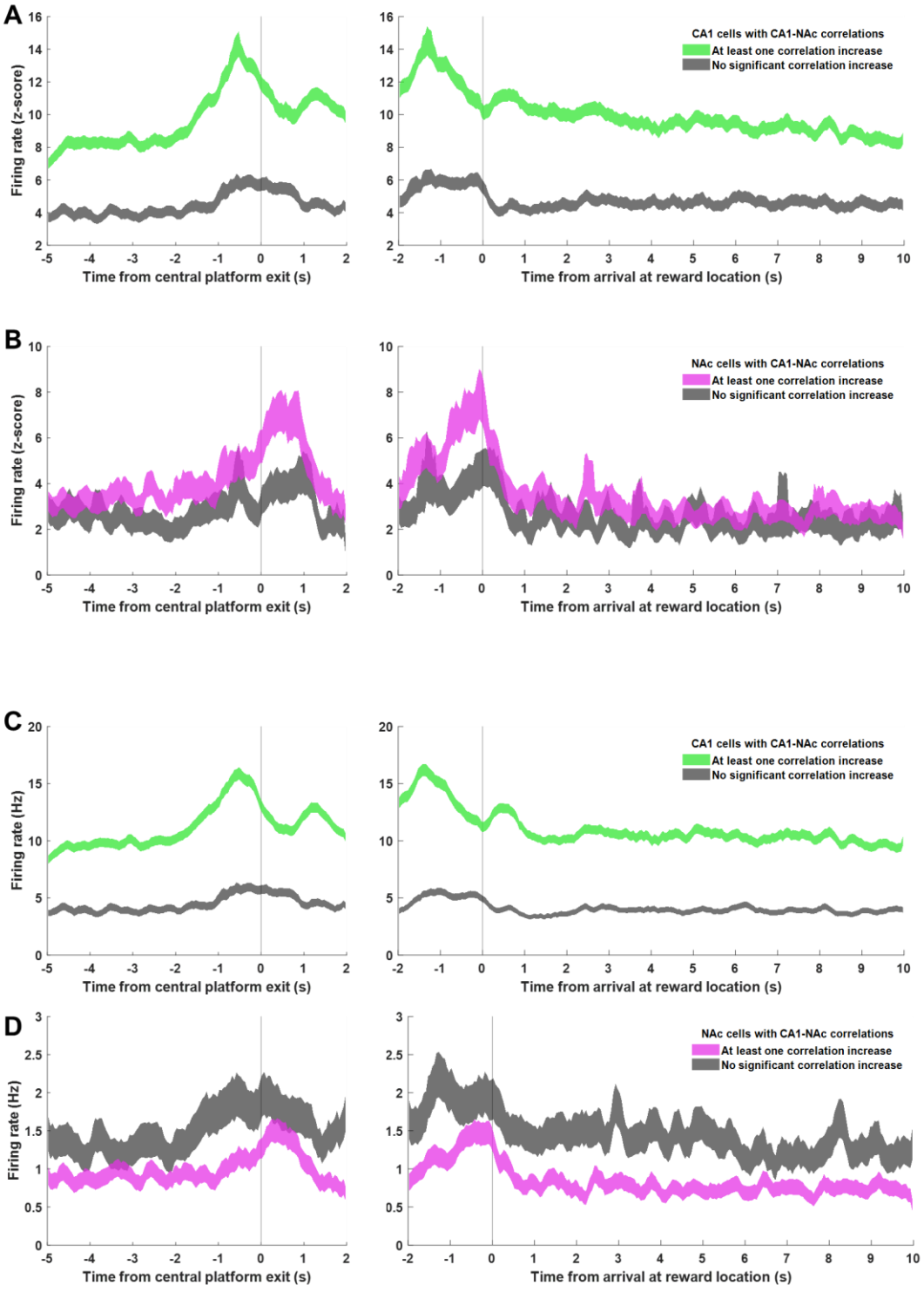

Figure 6.8. Firing rates of reactivated cells. Single-unit firing rates of the cells which comprise significantly reactivated CA1-NAc cells pairs, compared to the rest of CA1 and NAc cells which are not significantly reactivated with any cell. **A-B.** Z-scored firing rates of CA1 cells (A) and NAc cells (B), aligned to time from central platform exit (left) and time from arrival at reward location (right). **C-D.** As A-B, but with raw firing rates, not z-scored.

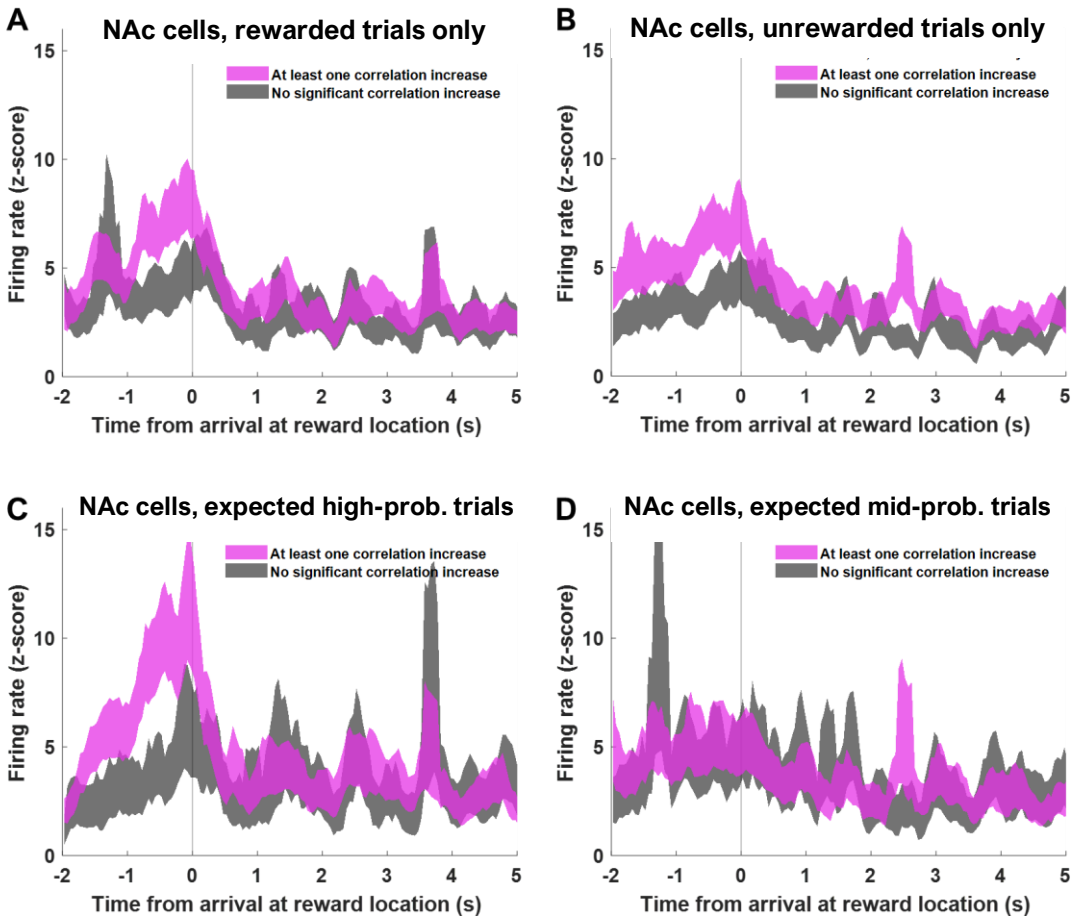

Figure 6.9. Firing rates of reactivated accumbens cells. Single-unit firing rates of the NAc cells which comprise significantly reactivated CA1-NAc cell pairs, compared to the rest of NAc cells which are not significantly reactivated with any cell. **A.** Firing rates on all rewarded trials, aligned to arrival at reward location. **B.** Firing rates on all unrewarded trials. **C-D.** Firing rates on rewarded trials in sessions 5-12; **C.** trials where the high-probability arm was chosen; **D.** trials where the mid-probability arm was chosen.

the lower overall firing rates of reactivated accumbens cells was apparent (fig. 6.8D). This ramping pattern has been widely reported in accumbens cells previously, and suggests a reward prediction signal (Khamassi et al, 2008; van der Meer & Redish, 2009) resulting from a ramping increase in dopamine signalling to the accumbens (Howe et al., 2013), which may be instructive for conveying reward-driven motivation. There was no difference in the theta-modulation index of reactivated versus non-reactivated accumbens cells (data not shown).

If the reactivated accumbens cells do encode a reward prediction signal, this ramping should be responsive to the expected reward but indifferent to reward outcome. Therefore, the firing rate of accumbens cells was compared for rewarded trials (fig. 6.9A) and unrewarded trials (fig 6.9B). The same ramping pattern was present on both trial types with little modulation of firing rates after reward outcome. Regarding reward expectancy, in sessions 5 to 12, where behavioural performance was above chance and before reward probabilities changed (Chapter 3), the rat could be said to have a good knowledge of the task demands and

reward probabilities, so reward expectancy was roughly accurate. Comparing activity on rewarded trials at the high-probability arm (which would elicit a small positive reward-prediction error) with activity on rewarded trials at the mid-probability arm (which would elicit a larger positive reward-prediction error), there was a much stronger effect of the former on firing rates (fig. 6.9C-D), indicating that reactivated accumbens cells preferentially encoded high reward-prediction. This joint activity of place information peaking at the start of the trajectory (coded by the CA1 cells) and reward prediction (accumbens) may be a way by which place and reward information is jointly reactivated, showing a higher correlation during POST than PRE.

6.3. Discussion

Pairs of CA1-accumbens cells were analysed for their correlated firing during performance on the probabilistic maze task and during rest periods before and after learning. In line with previous reports on replay in the accumbens and hippocampus, the majority of sessions showed significant variance in cell-pair correlations during post-task rest which could be explained by variance during task, an effect which persisted throughout the two-hour rest period.

A subset of CA1-accumbens cell pairs showed correlations during POST that were significantly greater than correlations during PRE. These cell pairs exhibited the strongest coactivity in their firing rate around the time of the trajectory towards the reward, a period in which there was also elevated theta-coherence in the local field potential of the two areas. This is consistent with previous reports that accumbens cells which encode reward-related or spatial information play a privileged role in hippocampus-mediated replay (Lansink et al., 2008; Lansink et al., 2009; Sjulson et al., 2018).

More specifically, the CA1 partners in these CA1-accumbens pairs showed a peak firing rate at the start of the trajectory towards reward, just prior to the exit from the central platform, while the accumbens partners showed a ramping up of activity towards the reward. Their joint reactivation during POST, which was stronger than during PRE, may be a way to bind place and reward prediction to inform future decisions when the animal is on the central platform and reinforce actions with high predicted reward.

Notably, neither the coactivity of these reactivated cell-pairs, nor their independent firing rates, showed evidence of modulation by reward outcome. The reward input aggregated over multiple trials used to calculate a reward prediction must arise from another source, either inputs from within the accumbens or arising from the VTA which sends reward-prediction error teaching signals to the accumbens.

Surprisingly, although the explained variance between hippocampus and accumbens cells during POST was significantly greater than during PRE during ripples, for almost all sessions, the contribution of each cell to this measure was not associated with any other firing property. Most notably, cells which were positively modulated by ripples did not contribute more to explained variance than cells which were negatively modulated or unmodulated. This was despite the well-established association between ripples and replay, and the use of ripple times to calculate explained variance for this analysis. It is possible that the bin width used for calculating explained variance, 50ms, is sufficient to capture only correlations between monosynaptic CA1-accumbens cell pairs, which are likely to form the minority of functionally connected ones (Trouche et al., 2019).

How might the joint reactivation of place and reward-prediction information arise during subsequent replay? The accumbens receives spatial information from CA1 (Sjulson et al., 2018), computes a reward prediction, and transmits this via the ventral pallidum to the VTA. The VTA, in turn, releases dopamine when disinhibited by the accumbens into target structures including both CA1 and accumbens (Floresco et al., 2003), where it modulates synaptic plasticity. Dopamine release in the accumbens during learning has been shown to modulate activity at cortico-striatal and limbic-striatal synapses both presynaptically and postsynaptically, influencing information transmission and plasticity processes as they arise from other brain areas. By inhibiting less active synapses, dopamine has been shown to effectively filter inputs to selectively reinforce the more active inputs (Pennartz et al., 1992; Bamford et al., 2004), which may be a mechanism for promoting hippocampal influence over accumbens activity in the presence of reward or the anticipation of reward, both of which trigger dopamine release. VTA also shows replay of reward-related information during sleep and rest (Valdés et al., 2015; Gomperts et al., 2015). The role of dopamine transmission in this system during sleep as a possible mechanism for learning is not clear: the bursting activity of VTA cells typical during waking activity has a non-linear additive effect on dopamine release compared to the same spiking at a steady rate, but bursting has been reported at a much lower rate during slow-wave sleep (Floresco et al., 2003), so it may not function as a transmitter of error signals to provide direct reinforcement learning. Nevertheless, dopamine manipulations in overnight memory experiments show it does have a function in sleep-dependent memory consolidation (Feld et al., 2014; Grogan et al., 2015; Asfestani et al., 2019).

The accumbens plays a role in mediating between hippocampus and VTA during wake, and participates in replay during sleep and rest. The results presented here show, for the first time, preferential engagement of reactivated pairs of hippocampus and accumbens cells in predictions of high reward probability over predictions of medium reward probability. This might form part of a teaching signal to the hippocampus during sleep and rest, which reinforces actions which have previously been rewarded. A direct link to reward-prediction error, and how this might bias replay, is not clear from these data, but the prediction error signal in accumbens might mediate dopamine release at both hippocampal synapses and hippocampo-accumbens synapses to influence plasticity.

Chapter 7: Discussion

In the work described in this thesis, a combination of behavioural, electrophysiological and computational studies was used to probe the influence of probabilistic rewards on spatial learning and decision-making. Specifically, the aim was to understand how hippocampal replay contributes to probabilistic reward learning, by influencing the activity of hippocampus and nucleus accumbens. This work builds on previous findings that correlated activity between hippocampus and accumbens increases during post-task rest, particularly during hippocampal sharp-wave ripples, and furthermore that activity relating to place and reward during the task is preferentially replayed (Lansink et al., 2008; Lansink et al., 209; Sjulson et al., 2018). The findings presented here add further detail to the picture of how hippocampus and accumbens contribute to probabilistic learning during wake and sleep.

7.1. Summary of principal findings

7.1.1. Replay can enhance learning

A Q-learning algorithm (Watkins, 1989) was used to run simulations of a task presented to rats, in which a maze with three arms which delivered stochastic rewards associated with different probabilities formed a partially-observable Markov decision process. Computational modelling showed that adding replay to a Q-learning model trained to alternate between two of three arms for stochastic rewards altered performance, in line with hypothesised functions about the influence of biological replay on spatial learning (Chapter 4). Prioritising replay on a trial-by-trial basis impaired learning by overfitting the model to a subset of unrepresentative trials, regardless of how the trials were prioritised. But grouping trials of a similar type together – in terms of their state-action pairs – allowed a suitable balance to be obtained between prioritising the most useful information to replay and ensuring representative samples. Replaying state-action pairs

according to the probability of reward increased the speed of learning, providing some benefit, but not the asymptotic performance. Replaying state-action pairs according to the average reward-prediction error, however, increased the asymptotic performance, allowing better learning overall.

Q-learning is a very simple model-free reinforcement algorithm which is unconstrained by the biological details of neurons, and so, arguably, enhanced replay in a Q-learning model would not necessarily translate to enhanced replay in a neurophysiological network. However, a body of literature has found that Q-learning predicts and correlates well with much of the spiking activity and BOLD signal of the brain when learning from reward. Furthermore, the behavioural output of the model replicated the patterns of rats' behaviour on the same task (Chapter 2), in particular the tendency to choose actions in proportion to their expected reward (probability-matching) and not according to a policy of optimal behaviour. The dynamics of learning in the Q-learning model can therefore be said to have some validity for explaining the underpinnings of rat behaviour.

7.1.2. Performance is influenced by offline reinforcement learning biased by reward-prediction error

Having validated both the behavioural task and the computational model, the model parameters were fit to the behaviour of the rats to see which kind of replay prioritisation (if any) best fitted their learning performance (Chapter 5). After training on the rats' own experience of states, actions and rewards, assuming additional updates to Q-values between sessions (equivalent to replay) altered the accuracy of the model's prediction of rats' behaviour. Assuming random replay or reward-biased replay made predictive accuracy worse, indicating that this is not a good explanation of how replay influences reinforcement learning. But assuming replay in which state-action pairs are sampled in accordance with the reward-prediction error elicited from them made predictive accuracy better than assuming no replay at all. Although not direct evidence of replay, this result is indicative of the mechanisms of offline memory consolidation and how it impacts reinforcement learning. Previous studies which have looked at the rewardedness of replay content have largely relied on associating cells with obvious tuning curves during behaviour with activity during subsequent rest. For example, this has shown that the greatest reactivation is with accumbens cells which show some degree of firing rate modulation around the time of reward (whether before or after, encoding reward expectancy, outcome or error), or ones which have a higher firing rate in the presence of cocaine reward than saline. These say little about the principles governing what gets replayed or the functions they might serve, where computational models can suggest how replay influences learning.

7.1.2. Hippocampus and accumbens engage in reward-related replay

Finally, single-unit recordings were made in vivo from hippocampus and accumbens during learning of a probabilistic maze task and during prior and subsequent rest. Single-unit activity and local field potential activity in both areas, as well as coherence between them, was modulated by the task (Chapter 3). In 15 sessions out of 17, the variation in correlations between firing rates of pairs of hippocampus-accumbens cells was more similar during post-task rest compared to during the task, than pre-task rest (Chapter 6). Corroborating previous findings, this indicates significant reactivation of task-related activity across the two brain regions (Lansink et al., 2018).

A limited dataset of just one rat means these results are preliminary and somewhat inconclusive, but the accumbens cells which showed the greatest increase in firing-rate correlations with hippocampus showed a ramping increase of their firing rate as the rat approached a reward location, and, moreover, this was selective for the high-probability arm when reward expectancy was high. Interestingly, this suggests replay of reward-prediction information, but there was no evidence of replay of a feedback or teaching signal (e.g. one which increased in response to the reward outcome to encode a reward-prediction error). Although this result would need more work (more data, and further analysis) to robustly link accumbens activity to the predictions of a reinforcement-learning algorithm such as the Q-learning one used here, this is a novel discovery and one which holds promise for delineating how offline activity in the accumbens relates to learning.

7.2. Discussion

The hippocampus and nucleus accumbens are widely reported to be involved in memory and learning. The hippocampus is crucial for spatial navigation, as demonstrated by countless patients with hippocampal lesions and non-human animals which undergo lesioning or inactivation of the hippocampus, who exhibit impaired spatial learning and memory subsequently. Disruption of accumbens activity results in impairments to learning from reward, especially when action selection involves a degree of ambiguity, uncertainty or risk. Changes in synaptic strength is apparent in both regions over the course of learning, demonstrating that they undergo experience-dependent consolidation which changes their synaptic responses, likely a substrate of changing behaviour.

Other brain regions are undoubtedly involved in the processing of both space (most notably entorhinal cortex) and reward (prefrontal cortex and VTA), and it is possible that neural representations of these features are distributed across networks in the brain rather than being confined to anatomical boundaries. In fact, even between the hippocampus and accumbens, the distribution of place and reward encoding is fuzzy: place cells have been found to be modulated by reward or distance from reward, while accumbens cells have been reported to encode as much spatial information as place cells in some circumstances (Sjulson et al., 2018). By experimental necessity, recordings in this project were limited to two brain regions, but many more are likely to take part in the encoding, learning and replay of task-related activity; these two brain regions do not exclusively hold the clues to reward-related replay. Nevertheless, their firing rates and coactivity during the task suggest that hippocampus and accumbens are both involved in this task.

The results from this work partly rely on a computational model which showed (a) that biasing replay by average reward-prediction error can boost learning, and (b) that rats behave as if they undergo this kind of replay between training sessions. Only one learning algorithm was selected to model the learning on this task. Common practice is to compare several models before selecting the most suitable, and indeed variations of Q-learning (Q-learning with forgetting, or Q-learning with separate learning rates for positive and negative reinforcement, for example) have been found in some cases to perform better than standard Q-learning in accounting for behaviour or neural activity. Such variations were not considered here, but could reveal more about how replay can influence learning if they were compared to standard Q-learning on this task. Other learning algorithms could also have proven to predict behaviour well: in particular the actor/critic framework, which has been widely used to model the basal ganglia and could account for the preferential replay of accumbens cells which predict reward, as seen in Chapter 6.

How might preferential replay between hippocampus and accumbens work? It has been suggested that the hippocampus generates predictions from its model of the world, which can be conveyed to other parts of the brain for the purposes of imagination, planning, and perhaps memory. This is likely to occur because

in its resting state (relatively unconstrained by sensorimotor input during behaviour) the organisation of connectivity in the hippocampus allows patterns of spikes to emerge (Scarpetta & Candia, 2013). These patterns originate in the CA3 subregion of the hippocampus, where highly recurrent, chaotic networks act as a pattern completer: a partial reinstatement of a memory trace (assembly activity of cells) can trigger the other cells in the assembly to reactivate as well, resulting in the full replay of the memory trace (Shen & McNaughton, 1996). The ease of reactivating the rest of the assembly depends on the functional connectivity between them, which is modulated by synaptic plasticity. Because only a partial reinstatement is necessary, noise in the network would be sufficient to occasionally activate enough cells at once from the same assembly for the pattern to be completed and a replay event to occur. This idea is supported by computational models which show that the theta activity observed during wakeful activity can trigger synaptic plasticity which then promotes the reactivation of the assembly activity during the period of high excitation that drives a sharp-wave ripple (Molter et al., 2007). Further computational work has shown that this plasticity during behaviour can be modulated by global reward, such that replay of activity which precedes a reward signal is preferentially replayed later (Miconi et al., 2017). Mesolimbic dopamine, widely considered to encode reward-related signals in much of the brain, is densely released in the accumbens and more sparsely in CA1, both modulating the excitability of cells in those areas during behaviour and promoting long-term plasticity withing and between them which persists for hours.

It is notable, for these reasons, that high theta coherence was observed between CA1 and accumbens during the approach to reward location, both in the results presented here and in previous reports (Lansink et al., 2016; van der Meer et al., 2019): perhaps entrainment to theta rhythms organises spike timing to augment plasticity between CA1 and accumbens, which preferentially engages the accumbens cells during subsequent hippocampal sharp-wave ripples. In agreement with this, accumbens cells which increased their firing rates around the time of hippocampal ripples showed greater theta-modulation of their firing rate than cells which decreased or did not change their firing rate around ripple times.

Finally, because of the chaotic firing patterns of the hippocampus at rest, small perturbations to the population activity can trigger replay events in the hippocampus which otherwise would not have occurred, allowing a mechanism for activity in other brain areas – most notably cortex and thalamus – to bias the overall replay content of the hippocampus, consistent with findings that sensory cues (where the flow of information is from sensory cortex to hippocampus) can bias replay (Bendor & Wilson, 2012; Schouten et al., 2017). Hence, it may only take small changes to excitability or synaptic potentiation during behaviour to have enduring effects of hippocampal replay, and for this to be conveyed throughout multiple brain areas to bias replay towards certain types of experience.

7.3. Future directions

Inevitably, this thesis does not resolve the question of what biases replay in the hippocampus-accumbens network, but it does extend the work that has previously been done in this area. Further work could better characterise how replay is coordinated across the brain in response to reward, with extended experimental studies to fully characterise the circuit involved in reinforcement learning during wake and sleep, and computational studies to generate and test theories about how replay can influence learning.

Replay has been observed in many other areas of the brain: prefrontal, parietal and sensory cortex, anterior cingulate cortex, amygdala, and VTA, any of which might replay activity synchronously with hippocampus and accumbens and influence the processing and reprocessing of task-related neural activity (Rothschild, 2019). In particular, reward-related replay has been found which is coordinated between hippocampus and accumbens (Lansink et al., 2009; Sjulson et al., 2018) and between hippocampus and VTA (Gomperts et al., 2015), and given the connectivity between them, there is good reason to hypothesise that accumbens and VTA might be reactivated together given the feedback loop that they form during wake. To date, this experiment has not been done, but it would elucidate reward-related processing and consolidation if replay across the whole circuit was characterised. In particular, given the apparent finding that VTA engages in hippocampal replay events during awake rest but not during sleep (Gomperts et al., 2015), and the globally low concentrations of dopamine during slow-wave sleep, such an experiment might elucidate the possible differences between replay during wake and replay during sleep, both in terms of cellular physiology and cognitive function.

Replay analysis typically depends on tracking changes in firing, or more commonly cell-pair correlations, over a period of minutes to hours within one recording session. On tasks such as the one central to this thesis, learning takes place over a period of days or weeks, so gradual changes in activity over that time course are not exposed. Electrophysiological methods make it difficult to identify the same cell in multiple recordings, but other techniques can be employed that might allow tracking of experience-dependent changes over hours to days or weeks as learning changes. This approach has been used in the hippocampus to refute existing theories about fear conditioning (Ahmed et al., 2019), and could also provide insight into the organisation of hippocampal and accumbens representations of reward-related activity as animals grow more familiar with a task or are faced with uncertainty.

Aside from the experiments, the Q-learning model employed in this thesis is rudimentary and leaves scope for further development. In particular, the model does not take into account activity at a neuronal level but provides only a mechanistic explanation: a more detailed computational model could allow investigation of how synaptic plasticity at hippocampal-accumbens synapses can change excitability of cells in the accumbens, driving different actions in response to the same spatial input, and furthermore how dopamine

release from VTA might interact with these processes (Humphries et al., 2009; Moyer et al., 2007). Dopamine has different effects on subsets of accumbens cells which primarily express D1 receptors or D2 receptors, resulting in somewhat parallel pathways, and in rats undergoing cocaine-conditioned place preference (a different kind of place-reward learning) the synapses between hippocampal cells and accumbens D2 cells were found to be selectively strengthened (Sjulson et al., 2018); they did not report whether D1 or D2 cells were preferentially engaged in hippocampal replay. Further modelling could uncover how some of these processes work.

7.4. Conclusion

The ability to learn from past experience to guide future behaviour is crucial for any animal or intelligent agent to navigate the world successfully. In nervous systems this is largely achieved with synaptic plasticity, through which the connections between neurons change their strength with the ultimate effect of altering the flow of neural activity from sensory input to processing areas of the brain and ultimately to motor output. Altering the policies for how to respond to the environment is a complex task, and the mechanisms by which it is achieved in the brain are subject to much investigation and debate.

What is apparent, though, is that some of this processing and reorganisation of neural activity occurs during sleep and rest. Activity during rest is dominated by different global dynamics in much of the brain, allowing it to effectively switch from processing ongoing events to past or future events, driven by internally generated activity more than externally. In the hippocampus, sharp-wave ripple events in the local field potential are the hallmark of such processing of temporally remote events, and coincide with activity in much of the rest of the brain, allowing consolidation of replayed memories at a systems level. This coordinated activity has an observable effect on subsequent behavioural policies, impairing spatial memory and reinforcement learning when disrupted.

The work in this thesis contributes to an ongoing picture that replay involves the recruitment of various disparate parts of the brain, not only to transfer recent memories from short-term to long-term storage, but to undergo additional processing as well. Here, the evidence suggests that replay emerges across the hippocampus and accumbens following learning which is biased towards certain experiences, and that this bias may boost learning.

Altogether, these findings highlight the complex dynamics that exist across behavioural states and sleep-wake cycles, and between brain regions. The topic of replay and memory consolidation has a long history, and as technology advances both in neuroscience (with the ability to manipulate neural activity with closed-loop stimulation, or to record from increasing large numbers of neurons) and machine learning (which shares a symbiotic relationship with neuroscience, mutually taking inspiration to generate new ideas), it is likely to remain a fruitful area for research.

References

Adam, S., Busoniu, L., & Babuska, R. (2012). Experience Replay for Real-Time Reinforcement Learning Control. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(2), 201–212. https://doi.org/10.1109/TSMCC.2011.2106494

Adcock, R. A., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B., & Gabrieli, J. D. E. (2006). Reward-Motivated Learning: Mesolimbic Activation Precedes Memory Formation. *Neuron*, *50*(3), 507–517. https://doi.org/10.1016/J.NEURON.2006.03.036

Ahmed, M. S., Priestley, J. B., Castro, A., Stefanini, F., Balough, E. M., Lavoie, E., … Losonczy, A. (2019). Hippocampal network reorganization underlies the formation of a temporal association memory. *BioRxiv*, 613638. https://doi.org/10.1101/613638

Aldridge, J. W., Berridge, K. C., & Rosen, A. R. (2004). Basal ganglia neural mechanisms of natural movement sequences. *Canadian Journal of Physiology and Pharmacology*, *82*(8–9), 732–739. https://doi.org/10.1139/y04-061

Ambrose, R. E., Pfeiffer, B. E., & Foster, D. J. (2016). Reverse Replay of Hippocampal Place Cells Is Uniquely Modulated by Changing Reward. *Neuron*, *91*(5), 1124–1136. https://doi.org/10.1016/j.neuron.2016.07.047

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., … Zaremba, W. (2017). *Advances in Neural Information Processing Systems 30 (NIPS 2017).* https://arxiv.org/pdf/1707.01495

Aronov, D., Nevers, R., & Tank, D. W. (2017). Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature*, *543*(7647), 719–722. https://doi.org/10.1038/nature21692

Asfestani, M. A., Brechtmann, V., Santiago, J., Born, J., & Feld, G. (2019). Consolidation of reward memory during sleep does not require dopaminergic activation. *BioRxiv*, 703132. https://doi.org/10.1101/703132

Aston-Jones, G., Rajkowski, J., Kubiak, P., & Alexinsky, T. (1994). Locus coeruleus neurons in monkey are selectively activated by attended cues in a vigilance task. *Journal of Neuroscience*, *14*(7), 4467–4480. https://doi.org/10.1523/JNEUROSCI.14-07-04467.1994

Atherton, L. A., Dupret, D., & Mellor, J. R. (2015). Memory trace replay: the shaping of memory consolidation by neuromodulation. *Trends in Neurosciences*, *38*(9), 560–570. https://doi.org/10.1016/j.tins.2015.07.004

Aubin, L., Khamassi, M., & Girard, B. (2018). *Prioritized Sweeping Neural DynaQ with Multiple Predecessors, and Hippocampal Replays*. https://doi.org/10.1007/978-3-319-95972-6_4

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral Prefrontal Cortex and Individual Differences in Uncertainty-Driven Exploration. *Neuron*, *73*(3), 595–607. https://doi.org/10.1016/J.NEURON.2011.12.025

Bamford, N. S., Zhang, H., Schmitz, Y., Wu, N. P., Cepeda, C., Levine, M. S., ... & Sulzer, D. (2004). Heterosynaptic dopamine neurotransmission selects sets of corticostriatal terminals. *Neuron*, *42*(4), 653-663. https://doi.org/10.1016/S0896-6273(04)00265-X

Battaglia, F. P., Sutherland, G. R., & McNaughton, B. L. (2004). Hippocampal sharp wave bursts coincide with neocortical "up-state" transitions. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *11*(6), 697–704. https://doi.org/10.1101/lm.73504

Beeler, J. A., Daw, N., Frazier, C. R. M., & Zhuang, X. (2010). Tonic Dopamine Modulates Exploitation of Reward Learning. *Frontiers in Behavioral Neuroscience*, *4*, 170. https://doi.org/10.3389/fnbeh.2010.00170

Behrend, E. R., & Bitterman, M. E. (1961). Probability-Matching in the Fish. *The American Journal of Psychology*, *74*(4), 542. https://doi.org/10.2307/1419664

Bendor, D., & Wilson, M. A. (2012). Biasing the content of hippocampal replay during sleep. *Nature Neuroscience*, *15*(10), 1439–1444. https://doi.org/10.1038/nn.3203

Berke, J. D., Okatan, M., Skurski, J., & Eichenbaum, H. B. (2004). Oscillatory entrainment of striatal neurons in freely moving rats. Neuron, 43(6), 883-896. https://doi.org/10.1016/j.neuron.2004.08.035

Berke, J. D. (2009). Fast oscillations in cortical-striatal networks switch frequency following rewarding events and stimulant drugs. *European Journal of Neuroscience*, *30*(5), 848–859. https://doi.org/10.1111/j.1460-9568.2009.06843.x

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. R. (2001). Predictability modulates human brain response to reward. *Journal of Neuroscience*, *21*(8), 2793–2798. https://doi.org/10.1523/jneurosci.6316-10.2011

Bertran-Gonzalez, J., Bosch, C., Maroteaux, M., Matamales, M., Hervé, D., Valjent, E., & Girault, J. A. (2008). Opposing patterns of signaling activation in dopamine D1 and D2 receptor-expressing striatal neurons in response to cocaine and haloperidol. *Journal of Neuroscience*, *28*(22), 5671-5685. https://doi.org/10.1523/JNEUROSCI.1039-08.2008

Bissonette, G. B., Burton, A. C., Gentry, R. N., Goldstein, B. L., Hearn, T. N., Barnett, B. R., … Roesch, M. R. (2013). Separate Populations of Neurons in Ventral Striatum Encode Value and Motivation. *PLoS ONE*, *8*(5), e64673. https://doi.org/10.1371/journal.pone.0064673

Bitterman, M., Wodinsky, J., & Candland, D. K. (1958). Some comparative psychology. *The American Journal of Psychology, 71(1)*, 94-110. https://doi.org/10.2307/1419199

Born, J., & Wilhelm, I. (2012). System consolidation of memory during sleep. *Psychological Research*, *76*(2), 192–203. https://doi.org/10.1007/s00426-011-0335-6

Bornstein, A. M., & Daw, N. D. (2011). Multiplicity of control in the basal ganglia: computational roles of striatal subregions. *Current Opinion in Neurobiology*, *21*(3), 374–380. https://doi.org/10.1016/J.CONB.2011.02.009

Bostock, E., Muller, R. U., & Kubie, J. L. (1991). Experience-dependent modifications of hippocampal place cell firing. *Hippocampus*, *1*(2), 193–205. https://doi.org/10.1002/hipo.450010207

Box, M., Jones, M. W., & Whiteley, N. (2016). A hidden Markov model for decoding and the analysis of replay in spike trains. *Journal of Computational Neuroscience*, *41*(3), 339–366. https://doi.org/10.1007/s10827-016-0621-9

Brimblecombe, K. R., & Cragg, S. J. (2017). The striosome and matrix compartments of the striatum: a path through the labyrinth from neurochemistry toward function. *ACS chemical neuroscience*, *8*(2), 235-242. https://doi.org/10.1021/acschemneuro.6b00333

Bruce, J., Suenderhauf, N., Mirowski, P., Hadsell, R., & Milford, M. (2017). *One-Shot Reinforcement Learning for Robot Navigation with Interactive Replay*. http://arxiv.org/abs/1711.10137

Buhry, L., Azizi, A. H., & Cheng, S. (2011). Reactivation, Replay, and Preplay: How It Might All Fit Together. *Neural Plasticity*, *2011*, 1–11. https://doi.org/10.1155/2011/203462

Bullock, D. H., & Bitterman, M. E. (1962). Probability-Matching in the Pigeon. *The American Journal of Psychology*, *75*(4), 634. https://doi.org/10.2307/1420288

Burton, A. C., Nakamura, K., & Roesch, M. R. (2015). From ventral-medial to dorsal-lateral striatum: Neural correlates of reward-guided decision-making. *Neurobiology of Learning and Memory*, *117*, 51–59. https://doi.org/10.1016/J.NLM.2014.05.003

Buzsáki, G. (2002). Theta Oscillations in the Hippocampus. *Neuron*, *33*(3), 325–340. https://doi.org/10.1016/S0896-6273(02)00586-X

Buzsáki, G. (2006). Rhythms of the Brain. Oxford University Press.

Buzsáki, G., & Vanderwolf, C. H. (1983). Cellular bases of hippocampal EEG in the behaving rat. *Brain Research Reviews*, *6*(2), 139-171. https://doi.org/10.1016/0165-0173(83)90037-1

Cachope, R., Mateo, Y., Mathur, B. N., Irving, J., Wang, H. L., Morales, M., ... & Cheer, J. F. (2012). Selective activation of cholinergic interneurons enhances accumbal phasic dopamine release: setting the tone for reward processing. *Cell reports*, *2*(1), 33-41. https://doi.org/10.1016/j.celrep.2012.05.011

Calabresi, P., Picconi, B., Tozzi, A., & di Filippo, M. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends in Neurosciences*, *30*(5), 211–219. https://doi.org/10.1016/J.TINS.2007.03.001

Cardinal, R. N., & Cheung, T. H. (2005). Nucleus accumbens core lesions retard instrumental learning and performance with delayed reinforcement in the rat. *BMC Neuroscience*, *6*(1), 9. https://doi.org/10.1186/1471-2202-6-9

Cardinal, R. N., & Howes, N. J. (2005). Effects of lesions of the nucleus accumbens core on choice between small certain rewards and large uncertain rewards in rats. *BMC Neuroscience*, *6*(1), 37. https://doi.org/10.1186/1471-2202-6-37

Carey, A. A., Tanaka, Y., & van der Meer, M. A. A. (2019). Reward revaluation biases hippocampal replay content away from the preferred outcome. *Nature Neuroscience*, *22*(9), 1450–1459. https://doi.org/10.1038/s41593-019-0464-6

Carmichael, J. E., Gmaz, J. M., & Meer, M. A. A. van der. (2017). Gamma Oscillations in the Rat Ventral Striatum Originate in the Piriform Cortex. *Journal of Neuroscience*, *37*(33), 7962–7974. https://doi.org/10.1523/JNEUROSCI.2944-15.2017

Carr, M. F., Jadhav, S. P., & Frank, L. M. (2011). Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nature Neuroscience*, *14*(2), 147–153. https://doi.org/10.1038/nn.2732

Carr, M. F., Karlsson, M. P., & Frank, L. M. (2012). Transient Slow Gamma Synchrony Underlies Hippocampal Memory Replay. *Neuron*, *75*(4), 700–713. https://doi.org/10.1016/J.NEURON.2012.06.014

Castillo Díaz, F., Hernandez, M. A., Capellá, T., & Medina, J. H. (2019). Dopamine Neurotransmission in the Ventral Tegmental Area Promotes Active Forgetting of Cocaine-Associated Memory. *Molecular Neurobiology*, *56*(9), 6206–6217. https://doi.org/10.1007/s12035-019-1516-3

Cazé, R., Khamassi, M., Aubin, L., & Girard, B. (2018). Hippocampal replays under the scrutiny of reinforcement learning models. *Journal of Neurophysiology*, *120*(6), 2877–2896. https://doi.org/10.1152/jn.00145.2018

Cembrowski, M. S., & Spruston, N. (2019). Heterogeneity within classical cell types is the rule: lessons from hippocampal pyramidal neurons. *Nature Reviews Neuroscience*, *20*(4), 193–204. https://doi.org/10.1038/s41583-019-0125-5

Chen, G., King, J. A., Burgess, N., & O'Keefe, J. (2013). How vision and movement combine in the hippocampal place code. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(1), 378–383. https://doi.org/10.1073/pnas.1215834110

Cheng, J., & Feenstra, M. G. P. (2006). Individual differences in dopamine efflux in nucleus accumbens shell and core during instrumental learning. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *13*(2), 168–177. https://doi.org/10.1101/lm.1806

Cheng, S., & Frank, L. M. (2008). New Experiences Enhance Coordinated Neural Activity in the Hippocampus. *Neuron*, *57*(2), 303–313. https://doi.org/10.1016/J.NEURON.2007.11.035

Chersi, F., & Pezzulo, G. (2012). Using hippocampal-striatal loops for spatial navigation and goal-directed decision-making. *Cognitive Processing*, *13*(S1), 125–129. https://doi.org/10.1007/s10339-012-0475-7

Chuhma, N., Mingote, S., Moore, H., & Rayport, S. (2014). Dopamine neurons control striatal cholinergic neurons via regionally heterogeneous dopamine and glutamate signaling. *Neuron*, *81*(4), 901-912. https://doi.org/10.1016/j.neuron.2013.12.027

Cichosz, P. (1999). An analysis of experience replay in temporal difference learning. *Cybernetics & Systems*, *30*(5), 341–363. https://doi.org/10.1080/019697299125127

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942. https://doi.org/10.1098/rstb.2007.2098

Colas, J. T., Pauli, W. M., Larsen, T., Tyszka, J. M., & O'Doherty, J. P. (2017). Distinct prediction errors in mesostriatal circuits of the human brain mediate learning about the values of both states and actions: evidence from high-resolution fMRI. *PLOS Computational Biology*, *13*(10), e1005810. https://doi.org/10.1371/journal.pcbi.1005810

Colgin, L. L., & Moser, E. I. (2010). Gamma Oscillations in the Hippocampus. *Physiology*, *25*(5), 319–329. https://doi.org/10.1152/physiol.00021.2010

Collins, A. L., Aitken, T. J., Huang, I. W., Shieh, C., Greenfield, V. Y., Monbouquette, H. G., ... & Wassum, K. M. (2019). Nucleus accumbens cholinergic interneurons oppose cue-motivated behavior. *Biological psychiatry*, *86*(5), 388-396. https://doi.org/10.1016/j.biopsych.2019.02.014

Cook, Z., Franks, D. W., & Robinson, E. J. H. (2013). Exploration versus exploitation in polydomous ant colonies. *Journal of Theoretical Biology*, *323*, 49–56. https://doi.org/10.1016/J.JTBI.2013.01.022

Corbit, L. H., Muir, J. L., & Balleine, B. W. (2001). The role of the nucleus accumbens in instrumental conditioning: Evidence of a functional dissociation between accumbens core and shell. *Journal of Neuroscience*, *21*(9), 3251–3260. https://doi.org/10.1523/JNEUROSCI.21-09-03251.2001

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, *319*(5867), 1264–1267. https://doi.org/10.1126/science.1150605

Dalton, G. L., Phillips, A. G., & Floresco, S. B. (2014). Preferential involvement by nucleus accumbens shell in mediating probabilistic learning and reversal shifts. *Journal of Neuroscience*, *34*(13), 4618–4626. https://doi.org/10.1523/JNEUROSCI.5058-13.2014

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

Daw, N. D., Niv, Y., & Dayan, P. (2005a). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. https://doi.org/10.1038/nn1560

de Lavilléon, G., Lacroix, M. M., Rondi-Reig, L., & Benchenane, K. (2015). Explicit memory creation during sleep demonstrates a causal role of place cells in navigation. *Nature Neuroscience*, *18*(4), 493–495. https://doi.org/10.1038/nn.3970

de Vivo, L., Bellesi, M., Marshall, W., Bushong, E. A., Ellisman, M. H., Tononi, G., & Cirelli, C. (2017). Ultrastructural evidence for synaptic scaling across the wake/sleep cycle. *Science*, *355*(6324), 507–510. https://doi.org/10.1126/science.aah5982

Delgado, M. R., Miller, M. M., Inati, S., & Phelps, E. A. (2005). An fMRI study of reward-related probability learning. *NeuroImage*, *24*(3), 862–873. https://doi.org/10.1016/J.NEUROIMAGE.2004.10.002

Deuker, L., Olligs, J., Fell, J., Kranz, T. A., Mormann, F., Montag, C., … Axmacher, N. (2013). Memory Consolidation by Replay of Stimulus-Specific Neural Activity. *Journal of Neuroscience*, *33*(49), 19373–19383. https://doi.org/10.1523/JNEUROSCI.0414-13.2013

Diba, K., & Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature Neuroscience*, *10*(10), 1241–1242. https://doi.org/10.1038/nn1961

Diekelmann, S., Wilhelm, I., & Born, J. (2009). The whats and whens of sleep-dependent memory consolidation. *Sleep Medicine Reviews*, *13*(5), 309–321. https://doi.org/10.1016/j.smrv.2008.08.002

Dragoi, G. (2013). Internal operations in the hippocampus: single cell and ensemble temporal coding. *Frontiers in systems neuroscience*, *7*, 46. https://doi.org/10.3389/fnsys.2013.00046

Dragoi, G., & Tonegawa, S. (2011). Preplay of future place cell sequences by hippocampal cellular assemblies. Nature, 469(7330), 397-401. https://doi.org/10.1038/nature09633

Dragoi, George, & Tonegawa, S. (2014). Selection of preconfigured cell assemblies for representation of novel spatial experiences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1635), 20120522. https://doi.org/10.1098/rstb.2012.0522

Dudai, Y. (2004). The Neurobiology of Consolidations, Or, How Stable is the Engram? *Annual Review of Psychology*, *55*(1), 51–86. https://doi.org/10.1146/annurev.psych.55.090902.142050

Dudek, S. M., Alexander, G. M., & Farris, S. (2016). Rediscovering area CA2: unique properties and functions. *Nature Reviews. Neuroscience*, *17*(2), 89–102. https://doi.org/10.1038/nrn.2015.22

Duszkiewicz, A. J., McNamara, C. G., Takeuchi, T., & Genzel, L. (2019). Novelty and Dopaminergic Modulation of Memory Persistence: A Tale of Two Systems. *Trends in Neurosciences*, *42*(2), 102–114. https://doi.org/10.1016/J.TINS.2018.10.002

Ego-Stengel, V., & Wilson, M. A. (2009). Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat. *Hippocampus*, *20*(1), NA-NA. https://doi.org/10.1002/hipo.20707

Eichenbaum, H. (2014). Time cells in the hippocampus: a new dimension for mapping memories. *Nature Reviews Neuroscience*, *15*(11), 732–744. https://doi.org/10.1038/nrn3827

Esmaeili, M. H., Kermani, M., Parvishan, A., & Haghparast, A. (2012). Role of D1/D2 dopamine receptors in the CA1 region of the rat hippocampus in the rewarding effects of morphine administered into the ventral tegmental area. *Behavioural brain research*, *231*(1), 111-115. https://doi.org/10.1016/j.bbr.2012.02.050

Exley, R., & Cragg, S. J. (2009). Presynaptic nicotinic receptors: a dynamic and diverse cholinergic filter of striatal dopamine neurotransmission. *British Journal of Pharmacology*, *153*(S1), S283–S297. https://doi.org/10.1038/sj.bjp.0707510

Feher da Silva, C., Victorino, C. G., Caticha, N., & Baldo, M. V. C. (2017). Exploration and recency as the main proximate causes of probability matching: a reinforcement learning analysis. *Scientific Reports*, *7*(1), 15326. https://doi.org/10.1038/s41598-017-15587-z

Feld, G. B., Besedovsky, L., Kaida, K., Münte, T. F., & Born, J. (2014). Dopamine D2-like Receptor Activation Wipes Out Preferential Consolidation of High over Low Reward Memories during Human Sleep. *Journal of Cognitive Neuroscience*, *26*(10), 2310–2320. https://doi.org/10.1162/jocn_a_00629

Fernández-Ruiz, A., Oliva, A., Fermino de Oliveira, E., Rocha-Almeida, F., Tingley, D., & Buzsáki, G. (2019). Long-duration hippocampal sharp wave ripples improve memory. *Science*, *364*(6445), 1082–1086. https://doi.org/10.1126/science.aax0758

Floresco, S. B. (2015). The Nucleus Accumbens: An Interface Between Cognition, Emotion, and Action. *Annual Review of Psychology*, *66*(1), 25–52. https://doi.org/10.1146/annurev-psych-010213-115159

Foster, D. J., & Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, *440*(7084), 680–683. https://doi.org/10.1038/nature04587

Foster, D. J., & Wilson, M. A. (2007). Hippocampal theta sequences. *Hippocampus*, *17*(11), 1093–1099. https://doi.org/10.1002/hipo.20345

Frank, L. M., Stanley, G. B., & Brown, E. N. (2004). Hippocampal Plasticity across Multiple Days of Exposure to Novel Environments. *Journal of Neuroscience*, *24*(35), 7681–7689. https://doi.org/10.1523/JNEUROSCI.1958-04.2004

Frankland, P. W., & Bontempi, B. (2005). The organization of recent and remote memories. *Nature Reviews Neuroscience*, *6*(2), 119–130. https://doi.org/10.1038/nrn1607

Franklin, N. T., & Frank, M. J. (2015). A cholinergic feedback circuit to regulate striatal population uncertainty and optimize reinforcement learning. *ELife*, *4*, e12029. https://doi.org/10.7554/eLife.12029

French, S. J., & Totterdell, S. (2002). Hippocampal and prefrontal cortical inputs monosynaptically converge with individual projection neurons of the nucleus accumbens. *Journal of Comparative Neurology*, *446*(2), 151-165. https://doi.org/10.1002/cne.10191

Frey, U., & Schroeder, H. (1990). Dopaminergic antagonists prevent long-term maintenance of posttetanic LTP in the CA1 region of rat hippocampal slices. *Brain research*, *522*(1), 69-75. https://doi.org/10.1016/0006-8993(90)91578-5

Fyhn, M., Molden, S., Witter, M. P., Moser, E. I., & Moser, M.-B. (2004). Spatial Representation in the Entorhinal Cortex. *Science*, *305*(5688), 1258–1264. https://doi.org/10.1126/SCIENCE.1099901

Frey, U., & Morris, R. G. M. (1998). Synaptic tagging: implications for late maintenance of hippocampal long-term potentiation. *Trends in Neurosciences*, *21*(5), 181–188. https://doi.org/10.1016/S0166-2236(97)01189-2

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*(3), 416–422. https://doi.org/10.1016/j.cognition.2008.09.007

Gagnon, D., Petryszyn, S., Sanchez, M. G., Bories, C., Beaulieu, J. M., De Koninck, Y., ... & Parent, M. (2017). Striatal neurons expressing D 1 and D 2 receptors are morphologically distinct and differently affected by dopamine denervation in mice. *Scientific reports*, *7*(1), 1-16. https://doi.org/ 10.1038/srep41432

Gangarossa, G., Espallergues, J., De Kerchove, D. E., El Mestikawy, S., Gerfen, C., Hervé, D., ... & Valjent, E. (2013). Distribution and compartmental organization of GABAergic medium-sized spiny neurons in the mouse nucleus accumbens. *Frontiers in neural circuits*, *7*, 22. https://doi.org/10.3389/fncir.2013.00022

Gardner, M. P. H., Schoenbaum, G., & Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1891), 20181645. https://doi.org/10.1098/rspb.2018.1645

Gauthier, J. L., & Tank, D. W. (2018). A Dedicated Population for Reward Coding in the Hippocampus. *Neuron*, *99*(1), 179-193.e7. https://doi.org/10.1016/J.NEURON.2018.06.008

Gerfen, C. R., Engber, T. M., Mahan, L. C., Susel, Z. V. I., Chase, T. N., Monsma, F. J., & Sibley, D. R. (1990). D1 and D2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. *Science*, *250*(4986), 1429-1432. https://doi.org/ 10.1126/science.2147780

Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., & Brea, J. (2018). Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. *Frontiers in Neural Circuits*, *12*, 53. https://doi.org/10.3389/FNCIR.2018.00053

Girardeau, G., Benchenane, K., Wiener, S. I., Buzsáki, G., & Zugaro, M. B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. *Nature Neuroscience*, *12*(10), 1222–1223. https://doi.org/10.1038/nn.2384

Girardeau, G., Inema, I., & Buzsáki, G. (2017). Reactivations of emotional memory in the hippocampus–amygdala system during sleep. *Nature Neuroscience*, *20*(11), 1634–1642. https://doi.org/10.1038/nn.4637

Giri, B., Miyawaki, H., Mizuseki, K., Cheng, S., & Diba, K. (2019). Hippocampal Reactivation Extends for Several Hours Following Novel Experience. *Journal of Neuroscience*, *39*(5), 866–875. https://doi.org/10.1523/JNEUROSCI.1950-18.2018

Gittins, J. C., & Jones, D. M. (1979). A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, *66*(3), 561–565. https://doi.org/10.1093/biomet/66.3.561

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, *66*(4), 585–595. https://doi.org/10.1016/J.NEURON.2010.04.016

Gmaz, J. M., Carmichael, J. E., & van der Meer, M. A. (2018). Persistent coding of outcome-predictive cue features in the rat nucleus accumbens. *ELife*, *7*. https://doi.org/10.7554/eLife.37275

Goldberg, J. A., & Reynolds, J. N. J. (2011). Spontaneous firing and evoked pauses in the tonically active cholinergic interneurons of the striatum. *Neuroscience*, *198*, 27–43. https://doi.org/10.1016/J.NEUROSCIENCE.2011.08.067

Gomperts, S. N., Kloosterman, F., Wilson, M. A., Cardinal, RN., Parkinson, JA., Hall, J., … Sejnowski, TJ. (2015). VTA neurons coordinate with the hippocampal reactivation of spatial experience. *ELife*, *4*, 321–352. https://doi.org/10.7554/eLife.05360

Gothard, K. M., Skaggs, W. E., & McNaughton, B. L. (1996). Dynamics of mismatch correction in the hippocampal ensemble code for space: interaction between path integration and environmental cues. *Journal of Neuroscience*, *16*(24), 8027–8040. https://doi.org/10.1523/JNEUROSCI.16-24-08027.1996

Graf, V., Bullock, D. H., & Bitterman, M. E. (1964). Further experiments on probability-matching in the pigeon1. *Journal of the Experimental Analysis of Behavior*, *7*(2), 151–157. https://doi.org/10.1901/jeab.1964.7-151

Graybiel, A. M. (1998). The Basal Ganglia and Chunking of Action Repertoires. *Neurobiology of Learning and Memory*, *70*(1–2), 119–136. https://doi.org/10.1006/NLME.1998.3843

Groenewegen, H. J., Wright, C. I., Beijer, A. V., & Voorn, P. (1999). Convergence and segregation of ventral striatal inputs and outputs. *Annals of the New York Academy of Sciences*, *877*(1), 49-63. https://doi.org/10.1111/j.1749-6632.1999.tb09260.x

Grogan, J., Bogacz, R., Tsivos, D., Whone, A., & Coulthard, E. (2015). Dopamine and Consolidation of Episodic Memory: Timing Is Everything. *Journal of Cognitive Neuroscience*, *27*(10), 2035–2050. https://doi.org/10.1162/jocn_a_00840

Gruber, M. J., Ritchey, M., Wang, S.-F., Doss, M. K., & Ranganath, C. (2016). Post-learning Hippocampal Dynamics Promote Preferential Retention of Rewarding Events. *Neuron*, *89*(5), 1110–1120. https://doi.org/10.1016/J.NEURON.2016.01.017

Gupta, A. S., van der Meer, M. A. A., Touretzky, D. S., & Redish, A. D. (2012). Segmentation of spatial experience by hippocampal θ sequences. *Nature Neuroscience*, *15*(7), 1032–1039. https://doi.org/10.1038/nn.3138

Haber, S. N., Fudge, J. L., & McFarland, N. R. (2000). Striatonigrostriatal Pathways in Primates Form an Ascending Spiral from the Shell to the Dorsolateral Striatum. *Journal of Neuroscience*, *20*(6), 2369–2382. https://doi.org/10.1523/JNEUROSCI.20-06-02369.2000

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801–806. https://doi.org/10.1038/nature03721

Hake, H. W., & Hyman, R. (1953). Perception of the statistical structure of a random series of binary symbols. *Journal of Experimental Psychology*, *45*(1), 64–74. https://doi.org/10.1037/h0060873

Han, P., Nakanishi, S. T., Tran, M. A., & Whelan, P. J. (2007). Dopaminergic modulation of spinal neuronal excitability. *Journal of Neuroscience*, *27*(48), 13192–13204. https://doi.org/10.1523/JNEUROSCI.1279-07.2007

Harris, K. D., Hirase, H., Leinekugel, X., Henze, D. A., & Buzsáki, G. (2001). Temporal interaction between single spikes and complex spike bursts in hippocampal pyramidal cells. *Neuron*, *32*(1), 141-149. https://doi.org/10.1016/S0896-6273(01)00447-0

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/J.NEURON.2017.06.011

Heimer, L., Zahm, D. S., Churchill, L., Kalivas, P. W., & Wohltmann, C. (1991). Specificity in the projection patterns of accumbal core and shell in the rat. *Neuroscience*, *41*(1), 89-125. https://doi.org/10.1016/0306-4522(91)90202-Y

Hernandez, P. J., Sadeghian, K., & Kelley, A. E. (2002). Early consolidation of instrumental learning requires protein synthesis in the nucleus accumbens. *Nature Neuroscience*, *5*(12), 1327–1331. https://doi.org/10.1038/nn973

Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., ... & Silver, D. (2018, April). Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Hirase, H., Leinekugel, X., Czurkó, A., Csicsvari, J., & Buzsáki, G. (2001). Firing rates of hippocampal neurons are preserved during subsequent sleep episodes and modified by novel awake experience. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(16), 9386–9390. https://doi.org/10.1073/pnas.161274398

Hollup, S. A., Kjelstrup, K. G., Hoff, J., Moser, M. B., & Moser, E. I. (2001). Impaired recognition of the goal location during spatial navigation in rats with hippocampal lesions. *Journal of Neuroscience*, *21*(12), 4505–4513. https://doi.org/10.1523/JNEUROSCI.21-12-04505.2001

Howe, M. W., Tierney, P. L., Sandberg, S. G., Phillips, P. E. M., & Graybiel, A. M. (2013). Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature*, *500*(7464), 575–579. https://doi.org/10.1038/nature12475

Hume, A. L., & Irwin, R. J. (1974). Bias functions and operating characteristics of rats discriminating auditory stimuli. *Journal of the Experimental Analysis of Behavior*, *21*(2), 1333196. https://doi.org/10.1901/jeab.1974.21-285

Humphries, M. D., & Prescott, T. J. (2010). The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward. *Progress in Neurobiology*, *90*(4), 385–417. https://doi.org/10.1016/J.PNEUROBIO.2009.11.003

Humphries, M. D., Wood, R., & Gurney, K. (2009a). Dopamine-modulated dynamic cell assemblies generated by the GABAergic striatal microcircuit. *Neural Networks*, *22*(8), 1174–1188. https://doi.org/10.1016/J.NEUNET.2009.07.018

Humphries, M. D., Wood, R., & Gurney, K. (2009b). Dopamine-modulated dynamic cell assemblies generated by the GABAergic striatal microcircuit. *Neural Networks*, *22*(8), 1174–1188. https://doi.org/10.1016/J.NEUNET.2009.07.018

Humphries, M., Khamassi, M., & Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in Neuroscience*, *6*, 9. https://doi.org/10.3389/fnins.2012.00009

Igloi, K., Gaggioni, G., Sterpenich, V., & Schwartz, S. (2015). A nap to recap or how reward regulates hippocampal-prefrontal memory networks during daytime sleep in humans. *Elife 4,* e07903. https://doi.org/10.7554/eLife.07903.001

Isele, D., & Cosgun, A. (2018). *Selective Experience Replay for Lifelong Learning*. http://arxiv.org/abs/1802.10269

Isotalus, H. K. (2019). Memory and The Ageing Brain: Dopamine, sleep and the hippocampus. *Ph. D. Thesis, University of Bristol*.

Ito, H. T., Zhang, S.-J., Witter, M. P., Moser, E. I., & Moser, M.-B. (2015). A prefrontal–thalamo–hippocampal circuit for goal-directed spatial navigation. *Nature*, *522*(7554), 50–55. https://doi.org/10.1038/nature14396

Ito, M., & Doya, K. (2009). Validation of Decision-Making Models and Analysis of Decision Variables in the Rat Basal Ganglia. *Journal of Neuroscience*, *29*(31), 9861–9874. https://doi.org/10.1523/jneurosci.6157-08.2009

Ito, Makoto, & Doya, K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current Opinion in Neurobiology*, *21*(3), 368–373. https://doi.org/10.1016/J.CONB.2011.04.001

Jackson, J. C., Johnson, A., & Redish, A. D. (2006). Hippocampal sharp waves and reactivation during awake states depend on repeated sequential experience. *Journal of Neuroscience*, *26*(48), 12415–12426. https://doi.org/10.1523/JNEUROSCI.4118-06.2006

Jadhav, S. P., Kemere, C., German, P. W., & Frank, L. M. (2012a). Awake Hippocampal Sharp-Wave Ripples Support Spatial Memory. *Science*, *336*(6087), 1454–1458. https://doi.org/10.1126/SCIENCE.1217230

Jang, A. I., Nassar, M. R., Dillon, D. G., & Frank, M. J. (2019). Positive reward prediction errors during decision-making strengthen memory encoding. *Nature human behaviour*, *3*(7), 719-732. https://doi.org/10.1038/s41562-019-0597-3

Jenkins, J. G., & Dallenbach, K. M. (1924). Obliviscence during sleep and waking. *The American Journal of Psychology*, *35*(4), 605-612. https://doi.org/ 10.2307/1414040

Jocham, G., Klein, T. A., & Ullsperger, M. (2011). Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. *Journal of Neuroscience*, *31*(5), 1606–1613. https://doi.org/10.1523/JNEUROSCI.3904-10.2011

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, *15*(4–6), 535–547. https://doi.org/10.1016/S0893-6080(02)00047-3

Johnson, A., & Redish, A. D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Networks*, *18*(9), 1163–1171. https://doi.org/10.1016/J.NEUNET.2005.08.009

Johnson, A., & Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, *27*(45), 12176–12189. https://doi.org/10.1523/JNEUROSCI.3761-07.2007

Joo, H. R., & Frank, L. M. (2018). The hippocampal sharp wave–ripple in memory retrieval for immediate use and consolidation. *Nature Reviews Neuroscience*, *19*(12), 744–757. https://doi.org/10.1038/s41583-018-0077-1

Jun, J. J., Steinmetz, N. A., Siegle, J. H., Denman, D. J., Bauza, M., Barbarits, B., … Harris, T. D. (2017). Fully integrated silicon probes for high-density recording of neural activity. *Nature*, *551*(7679), 232–236. https://doi.org/10.1038/nature24636

Kalyanakrishnan, S., & Stone, P. (2007). Batch reinforcement learning in a complex domain. *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems - AAMAS '07*, 1. https://doi.org/10.1145/1329125.1329241

Karimpanal, T. G., & Bouffanais, R. (2017). Experience Replay Using Transition Sequences. *Frontiers in Neurorobotics*, *12*, 32. https://doi.org/10.3389/fnbot.2018.00032

Karlsson, M., & Frank, L. (2009). Awake replay of remote experiences in the hippocampus. *Nature Neuroscience, 12(7),* 913-918. https://doi.org/10.1038/nn.2344

Kay, K., & Frank, L. M. (2019). Three brain states in the hippocampus and cortex. *Hippocampus*, *29*(3), 184–238. https://doi.org/10.1002/hipo.22956

Kelley, A. E., Smith-Roe, S. L., & Holahan, M. R. (1997). Response-reinforcement learning is dependent on N-methyl-D-aspartate receptor activation in the nucleus accumbens core. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(22), 12174–12179. https://doi.org/10.1073/pnas.94.22.12174

Kentros, C. G., Agnihotri, N. T., Streater, S., Hawkins, R. D., & Kandel, E. R. (2004). Increased Attention to Spatial Context Increases Both Place Field Stability and Spatial Memory. *Neuron*, *42*(2), 283–295. https://doi.org/10.1016/S0896-6273(04)00192-8

Kesner, R. P. (2007). Behavioral functions of the CA3 subregion of the hippocampus. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *14*(11), 771–781. https://doi.org/10.1101/lm.688207

Khamassi, M., & Humphries, M. D. (2012). Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies. *Frontiers in Behavioral Neuroscience*, *6*, 79. https://doi.org/10.3389/fnbeh.2012.00079

Khamassi, M., Mulder, A. B., Tabuchi, E., Douchamps, V., & Wiener, S. I. (2008). Anticipatory reward signals in ventral striatal neurons of behaving rats. *European Journal of Neuroscience*, *28*(9), 1849–1866. https://doi.org/10.1111/j.1460-9568.2008.06480.x

Kim, H., Lee, D., & Jung, M. W. (2013). Signals for Previous Goal Choice Persist in the Dorsomedial, but Not Dorsolateral Striatum of Rats. *Journal of Neuroscience*, *33*(1), 52–63. https://doi.org/10.1523/jneurosci.2422-12.2013

Kim, Hoseok, Sul, J. H., Huh, N., Lee, D., & Jung, M. W. (2009). Role of striatum in updating values of chosen actions. *Journal of Neuroscience*, *29*(47), 14701–14712. https://doi.org/10.1523/JNEUROSCI.2728-09.2009

King, C., Henze, D. A., Leinekugel, X., & Buzsáki, G. (1999). Hebbian modification of a hippocampal population pattern in the rat. *The Journal of Physiology*, *521*(1), 159–167. https://doi.org/10.1111/j.1469-7793.1999.00159.x

Kirk, K. L., & Bitterman, M. E. (1965). Probability-learning by the turtle. *Science*, *148*(3676), 1484–1485. https://doi.org/10.1126/science.148.3676.1484

Kitamura, T., Ogawa, S. K., Roy, D. S., Okuyama, T., Morrissey, M. D., Smith, L. M., … Tonegawa, S. (2017). Engrams and circuits crucial for systems consolidation of a memory. *Science*, *356*(6333), 73–78. https://doi.org/10.1126/science.aam6808

Kloosterman, F., Layton, S. P., Chen, Z., & Wilson, M. A. (2014). Bayesian decoding using unsorted spikes in the rat hippocampus. *Journal of Neurophysiology*, *111*(1), 217–227. https://doi.org/10.1152/jn.01046.2012

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience*, *21*(16), RC159. https://doi.org/10.1523/JNEUROSCI.21-16-j0002.2001

Kobayashi, T., Hori, E., Umeno, K., Tazumi, T., Ono, T., & Nishijo, H. (2006). Conjunctive effects of reward and behavioral episodes on hippocampal place-differential neurons of rats on a mobile treadmill. *Hippocampus*, *16*(7), 586–595. https://doi.org/10.1002/hipo.20186

Koehler, D. J., & James, G. (2014). Probability Matching, Fast and Slow. *Psychology of Learning and Motivation*, *61*, 103–131. https://doi.org/10.1016/B978-0-12-800283-4.00003-4

Komorowski, R. W., Manns, J. R., & Eichenbaum, H. (2009). Robust conjunctive item-place coding by hippocampal neurons parallels learning what happens where. *Journal of Neuroscience*, *29*(31), 9918–9929. https://doi.org/10.1523/JNEUROSCI.1378-09.2009

Kreitzer, A. C. (2009). Physiology and pharmacology of striatal neurons. *Annual review of neuroscience*, *32*, 127-147. https://doi.org/10.1146/annurev.neuro.051508.135422

Kruskal, P. B., Stanis, J. J., McNaughton, B. L., & Thomas, P. J. (2007). A binless correlation measure reduces the variability of memory reactivation estimates. *Statistics in Medicine*, *26*(21), 3997–4008. https://doi.org/10.1002/sim.2946

Kudrimoti, H. S., Barnes, C. A., & McNaughton, B. L. (1999). Reactivation of hippocampal cell assemblies: effects of behavioral state, experience, and EEG dynamics. *Journal of Neuroscience*, *19*(10), 4090–4101. https://doi.org/10.1523/JNEUROSCI.19-10-04090.1999

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated. *Trends in Cognitive Sciences*, *20*(7), 512–534. https://doi.org/10.1016/J.TICS.2016.05.004

Kupchik, Y. M., Brown, R. M., Heinsbroek, J. A., Lobo, M. K., Schwartz, D. J., & Kalivas, P. W. (2015). Coding the direct/indirect pathways by D1 and D2 receptors is not valid for accumbens projections. *Nature neuroscience*, *18*(9), 1230-1232. https://doi.org/10.1038/nn.4068

Lalonde, R. (2002). The neurobiological basis of spontaneous alternation. *Neuroscience & Biobehavioral Reviews*, *26*(1), 91–104. https://doi.org/10.1016/S0149-7634(01)00041-0

Lansink, C. S., Goltstein, P. M., Lankelma, J. v., Joosten, R. N. J. M. A., McNaughton, B. L., & Pennartz, C. M. A. (2008). Preferential Reactivation of Motivationally Relevant Information in the Ventral Striatum. *Journal of Neuroscience*, *28*(25). https://doi.org/10.1523/JNEUROSCI.1054-08.2008

Lansink, C. S., Goltstein, P. M., Lankelma, J. v., McNaughton, B. L., & Pennartz, C. M. A. (2009). Hippocampus Leads Ventral Striatum in Replay of Place-Reward Information. *PLoS Biology*, *7*(8), e1000173. https://doi.org/10.1371/journal.pbio.1000173

Lansink, C. S., Meijer, G. T., Lankelma, J. v, Vinck, M. A., Jackson, J. C., & Pennartz, C. M. A. (2016). Reward Expectancy Strengthens CA1 Theta and Beta Band Synchronization and Hippocampal-Ventral Striatal Coupling. *Journal of Neuroscience*, *36*(41), 10598–10610. https://doi.org/10.1523/JNEUROSCI.0682-16.2016

Lee, A. K., & Wilson, M. A. (2002). Memory of Sequential Experience in the Hippocampus during Slow Wave Sleep. *Neuron*, *36*(6), 1183–1194. https://doi.org/10.1016/S0896-6273(02)01096-6

Li, J., & Daw, N. D. (2011). Signals in human striatum are appropriate for policy update rather than value prediction. *Journal of Neuroscience*, *31*(14), 5504–5511. https://doi.org/10.1523/JNEUROSCI.6316-10.2011

Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, *8*(3–4), 293–321. https://doi.org/10.1007/BF00992699

Leutgeb, S., Ragozzino, K. E., & Mizumori, S. J. Y. (2000). Convergence of head direction and place information in the CA1 region of hippocampus. *Neuroscience*, *100*(1), 11–19. https://doi.org/10.1016/S0306-4522(00)00258-X

Lever, C., Wills, T., Cacucci, F., Burgess, N., & O'Keefe, J. (2002). Long-term plasticity in hippocampal place-cell representation of environmental geometry. *Nature*, *416*(6876), 90–94. https://doi.org/10.1038/416090a

Lewis, P. A., Knoblich, G., & Poe, G. (2018). How Memory Replay in Sleep Boosts Creative Problem-Solving. *Trends in Cognitive Sciences*, *22*(6), 491–503. https://doi.org/10.1016/J.TICS.2018.03.009

Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, *8*(3–4), 293–321. https://doi.org/10.1007/BF00992699

Lisman, J. E., & Grace, A. A. (2005). Review The Hippocampal-VTA Loop: Controlling the Entry of Information into Long-Term Memory The Role of the Hippocampus in Producing Novelty-Dependent Firing of VTA Cells Recordings from dopaminergic cells in awake monkeys. *Neuron*, *46*, 703–713. https://doi.org/10.1016/j.neuron.2005.05.002

Liu, R., & Zou, J. (2018). The Effects of Memory Replay in Reinforcement Learning. *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 478–485. https://doi.org/10.1109/ALLERTON.2018.8636075

Lopes-dos-Santos, V., Ribeiro, S., & Tort, A. B. L. (2013). Detecting cell assemblies in large neuronal populations. *Journal of Neuroscience Methods*, *220*(2), 149–166. https://doi.org/10.1016/J.JNEUMETH.2013.04.010

Lopes-dos-Santos, V., van de Ven, G. M., Morley, A., Trouche, S., Campo-Urriza, N., & Dupret, D. (2018). Parsing Hippocampal Theta Oscillations by Nested Spectral Components during Spatial Exploration and Memory-Guided Behavior. *Neuron*, *100*(4), 940-952.e7. https://doi.org/10.1016/J.NEURON.2018.09.031

Louie, K., & Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, *29*(1), 145–156. https://doi.org/10.1016/S0896-6273(01)00186-6

MacAskill, A. F., Cassel, J. M., & Carter, A. G. (2014). Cocaine exposure reorganizes cell type–and input-specific connectivity in the nucleus accumbens. *Nature neuroscience*, *17*(9), 1198-1207. https://doi.org/10.1038/nn.3783

MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal "Time Cells" Bridge the Gap in Memory for Discontiguous Events. *Neuron*, *71*(4), 737–749. https://doi.org/10.1016/J.NEURON.2011.07.012

Mahon, S., Vautrelle, N., Pezard, L., Slaght, S. J., Deniau, J. M., Chouvet, G., & Charpier, S. (2006). Distinct patterns of striatal medium spiny neuron activity during the natural sleep–wake cycle. *Journal of Neuroscience*, *26*(48), 12587-12595. https://doi.org/10.1523/JNEUROSCI.3987-06.2006

Malhotra, S., Cross, R. W., Zhang, A., & van der Meer, M. A. A. (2015). Ventral striatal gamma oscillations are highly variable from trial to trial, and are dominated by behavioural state, and only weakly influenced by outcome value. *European Journal of Neuroscience*, *42*(10), 2818–2832. https://doi.org/10.1111/ejn.13069

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an Integration of Deep Learning and Neuroscience. *Frontiers in Computational Neuroscience*, *10*, 94. https://doi.org/10.3389/fncom.2016.00094

Marshall, L., & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in Cognitive Sciences*, *11*(10), 442–450. https://doi.org/10.1016/J.TICS.2007.09.001

Martin, K. C., & Kosik, K. S. (2002). Synaptic tagging — who's it? *Nature Reviews Neuroscience*, *3*(10), 813–820. https://doi.org/10.1038/nrn942

Martin, S. J., de Hoz, L., & Morris, R. G. M. (2005). Retrograde amnesia: neither partial nor complete hippocampal lesions in rats result in preferential sparing of remote spatial memory, even after reminding. *Neuropsychologia*, *43*(4), 609–624. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2004.07.007

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, *21*(11), 1609–1617. https://doi.org/10.1038/s41593-018-0232-z

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal Prediction Errors in a Passive Learning Task Activate Human Striatum. *Neuron*, *38*(2), 339–346. https://doi.org/10.1016/S0896-6273(03)00154-5

McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., & Schoenbaum, G. (2011). Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *Journal of Neuroscience*, *31*(7), 2700–2705. https://doi.org/10.1523/JNEUROSCI.5499-10.2011

McNamara, C. G., & Dupret, D. (2017). Two sources of dopamine for the hippocampus. *Trends in Neurosciences*, *40*(7), 383–384. https://doi.org/10.1016/J.TINS.2017.05.005

McNamara, C. G., Tejero-Cantero, Á., Trouche, S., Campo-Urriza, N., & Dupret, D. (2014). Dopaminergic neurons promote hippocampal reactivation and spatial memory persistence. *Nature Neuroscience*, *17*(12), 1658–1660. https://doi.org/10.1038/nn.3843

McNaughton, B. L., Leonard, B., & Chen, L. (1989). Cortical-hippocampal interactions and cognitive mapping: A hypothesis based on reintegration of the parietal and inferotemporal pathways for visual processing. *Psychobiology*, *17*(3), 230–235. https://doi.org/10.1007/BF03337774

Meer, M. A. A. van der, Gmaz, J. M., & Carmichael, J. E. (2019). A comprehensive characterization of rhythmic spiking activity in the rat ventral striatum. *BioRxiv*, 617233. https://doi.org/10.1101/617233

Michon, F., Sun, J.-J., Kim, C. Y., Ciliberti, D., & Kloosterman, F. (2019). Post-learning Hippocampal Replay Selectively Reinforces Spatial Memory for Highly Rewarded Locations. *Current Biology*, *29*(9), 1436-1444.e5. https://doi.org/10.1016/J.CUB.2019.03.048

Miconi, T. (2017). Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *ELife*, *6*. https://doi.org/10.7554/eLife.20899

Miller, J. D., Farber, J., Gatz, P., Roffwarg, H., & German, D. C. (1983). Activity of mesencephalic dopamine and non-dopamine neurons across stages of sleep and waking in the rat. *Brain research*, *273*(1), 133-141. https://doi.org/10.1016/0006-8993(83)91101-0

Minsky, M. (1961). Steps toward Artificial Intelligence. *Proceedings of the IRE*, *49*(1), 8–30. https://doi.org/10.1109/JRPROC.1961.287775

Mishra, R. K., Kim, S., Guzman, S. J., & Jonas, P. (2016). Symmetric spike timing-dependent plasticity at CA3–CA3 synapses optimizes storage and recall in autoassociative networks. *Nature Communications*, *7*(1), 11552. https://doi.org/10.1038/ncomms11552

Miyawaki, T., Norimoto, H., Ishikawa, T., Watanabe, Y., Matsuki, N., & Ikegaya, Y. (2014). Dopamine Receptor Activation Reorganizes Neuronal Ensembles during Hippocampal Sharp Waves In Vitro. *PLoS ONE*, *9*(8), e104438. https://doi.org/10.1371/journal.pone.0104438

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). *Playing Atari with Deep Reinforcement Learning*. http://arxiv.org/abs/1312.5602

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

Mogenson, G. J., Jones, D. L., & Yim, C. Y. (1980). From motivation to action: Functional interface between the limbic system and the motor system. *Progress in Neurobiology*, *14*(2–3), 69–97. https://doi.org/10.1016/0301-0082(80)90018-0

Molter, C., Sato, N., & Yamaguchi, Y. (2007). Reactivation of behavioral activity during sharp waves: A computational model for two stage hippocampal dynamics. *Hippocampus*, *17*(3), 201–209. https://doi.org/10.1002/hipo.20258

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *ELife*, *7*. https://doi.org/10.7554/eLife.32548

Montague, P. R., Dayan, P., Sejnowski, T. J., & O'Doherty, J. P. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*(5), 1936–1947. https://doi.org/10.1523/jneurosci.2496-07.2007

Moore, A. W., & Atkeson, C. G. (1993). Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning*, *13*(1), 103–130. https://doi.org/10.1007/BF00993104

Morris, G., Schmidt, R., & Bergman, H. (2010). Striatal action-learning based on dopamine concentration. *Experimental Brain Research*, *200*(3–4), 307–317. https://doi.org/10.1007/s00221-009-2060-6

Moyer, J. T., Wolf, J. A., & Finkel, L. H. (2007). Effects of Dopaminergic Modulation on the Integrative Properties of the Ventral Striatal Medium Spiny Neuron. *Journal of Neurophysiology*, *98*(6), 3731–3748. https://doi.org/10.1152/jn.00335.2007

Muenzinger, K. F. (1931). The primary factors in learning. *Psychological Review*, *38*(4), 347–358. https://doi.org/10.1037/h0074319

Murphy, A. H., & Murphy, A. H. (1973). A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*, *12*(4), 595–600. https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2

Nádasdy, Z., Hirase, H., Czurkó, A., Csicsvari, J., & Buzsáki, G. (1999). Replay and time compression of recurring spike sequences in the hippocampus. *Journal of Neuroscience*, *19*(21), 9497–9507. https://doi.org/10.1523/JNEUROSCI.19-21-09497.1999

Narasimhan, K., Yala, A., & Barzilay, R. (2016). *Improving Information Extraction by Acquiring External Evidence with Reinforcement Learning*. http://arxiv.org/abs/1603.07954

Nguyen, D., Fugariu, V., Erb, S., & Ito, R. (2018). Dissociable roles of the nucleus accumbens D1 and D2 receptors in regulating cue-elicited approach-avoidance conflict decision-making. *Psychopharmacology*, *235*(8), 2233-2244. https://doi.org/10.1007/s00213-018-4919-3

Nicola, S. M. (2010). The flexible approach hypothesis: unification of effort and cue-responding hypotheses for the role of nucleus accumbens dopamine in the activation of reward-seeking behavior. *Journal of Neuroscience*, *30*(49), 16585–16600. https://doi.org/10.1523/JNEUROSCI.3958-10.2010

Nicola, S. M. (2007). The nucleus accumbens as part of a basal ganglia action selection circuit. *Psychopharmacology*, 191(3), 521-550. https://doi.org/10.1007/s00213-006-0510-4

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452–454. https://doi.org/10.1126/science.1094285

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron*, *38*(2), 329–337. https://doi.org/10.1016/S0896-6273(03)00169-7

O'Donnell, P., & Grace, A. A. (1996). Dopaminergic Reduction of Excitability in Nucleus Accumbens Neurons Recorded in Vitro. *Neuropsychopharmacology*, *15*(1), 87–97. https://doi.org/10.1016/0893-133X(95)00177-F

O'Keefe, J., & 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. https://doi.org/10.1016/0006-8993(71)90358-1

O'Keefe, John, & Recce, M. L. (1993). Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus*, *3*(3), 317–330. https://doi.org/10.1002/hipo.450030307

Ólafsdóttir, H. F., Barry, C., Saleem, A. B., Hassabis, D., & Spiers, H. J. (2015). Hippocampal place cells construct reward related sequences through unexplored space. *ELife*, *4*. https://doi.org/10.7554/eLife.06063

Ólafsdóttir, H. F., Bush, D., & Barry, C. (2018). The Role of Hippocampal Replay in Memory and Planning. *Current Biology*, *28*(1), R37–R50. https://doi.org/10.1016/J.CUB.2017.10.073

Ólafsdóttir, H. F., Carpenter, F., & Barry, C. (2017a). Task Demands Predict a Dynamic Switch in the Content of Awake Hippocampal Replay. *Neuron*, *96*(4), 925-935.e6. https://doi.org/10.1016/j.neuron.2017.09.035

O'Neill, J., Senior, T. J., Allen, K., Huxter, J. R., & Csicsvari, J. (2008). Reactivation of experience-dependent cell assembly patterns in the hippocampus. *Nature Neuroscience*, *11*(2), 209–215. https://doi.org/10.1038/nn2037

Otmakhova, N. A., & Lisman, J. E. (1996). D1/D5 dopamine receptor activation increases the magnitude of early long-term potentiation at CA1 hippocampal synapses. *Journal of Neuroscience*, *16*(23), 7478-7486. https://doi.org/10.1523/JNEUROSCI.16-23-07478.1996

Otto, A. R., Taylor, E. G., & Markman, A. B. (2011). There are at least two kinds of probability matching: Evidence from a secondary task. *Cognition*, *118*(2), 274–279. https://doi.org/10.1016/J.COGNITION.2010.11.009

Oudiette, D., & Paller, K. A. (2013). Upgrading the sleeping brain with targeted memory reactivation. *Trends in Cognitive Sciences*, *17*(3), 142–149. https://doi.org/10.1016/J.TICS.2013.01.006

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, *5*(2), 97–98. https://doi.org/10.1038/nn802

Papageorgiou, G. K., Baudonnat, M., Cucca, F., & Walton, M. E. (2016). Mesolimbic dopamine encodes prediction errors in a state-dependent manner. *Cell reports*, *15*(2), 221-228. https://doi.org/10.1016/j.celrep.2016.03.031

Pavlides, C., & Winson, J. (1989). Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *Journal of Neuroscience*, *9*(8), 2907–2918. https://doi.org/10.1523/JNEUROSCI.09-08-02907.1989

Paxinos, G., & Watson, C. (1996). The Rat Brain Atlas in Stereotaxic Co-ordinates 4th edn. *Academic, Sydney*.

Pennartz, Ito, Verschure, Battaglia, & Robbins. (2011). The hippocampal-striatal axis in learning, prediction and goal-directed behavior. *Trends in Neurosciences*, *34*(10). https://doi.org/10.1016/j.tins.2011.08.001

Pennartz, C. M. A., Lee, E., Verheul, J., Lipa, P., Barnes, C. A., & McNaughton, B. L. (2004). The Ventral Striatum in Off-Line Processing: Ensemble Reactivation during Sleep and Modulation by Hippocampal Ripples. *Journal of Neuroscience*, *24*(29), 6446–6456. https://doi.org/10.1523/JNEUROSCI.0575-04.2004

Peyrache, A., Khamassi, M., & Benchenane, K. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nature Neuroscience, 12,* 919-926. https://doi.org/10.1038/nn.2337

Pezzulo, G., Kemere, C., & van der Meer, M. A. A. (2017). Internally generated hippocampal sequences as a vantage point to probe future-oriented cognition. *Annals of the New York Academy of Sciences*, *1396*(1), 144–165. https://doi.org/10.1111/nyas.13329

Plenz, D. (2003). When inhibition goes incognito: feedback interaction between spiny projection neurons in striatal function. *Trends in neurosciences*, *26*(8), 436-443. https://doi.org/10.1016/S0166-2236(03)00196-6

Plonsky, O., Teodorescu, K., & Erev, I. (2015). Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychological Review*, *122*(4), 621–647. https://doi.org/10.1037/a0039413

Preuschoff, K., Bossaerts, P., & Quartz, S. R. (2006). Neural Differentiation of Expected Reward and Risk in Human Subcortical Structures. *Neuron*, *51*(3), 381–390. https://doi.org/10.1016/J.NEURON.2006.06.024

Qin, Y.-L., Mcnaughton, B. L., Skaggs, W. E., & Barnes, C. A. (1997). Memory reprocessing in corticocortical and hippocampocortical neuronal ensembles. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *352*(1360), 1525–1533. https://doi.org/10.1098/rstb.1997.0139

Quirk, M. C., & Wilson, M. A. (1999). Interaction between spike waveform classification and temporal sequence detection. *Journal of neuroscience methods*, *94*(1), 41-52. https://doi.org/10.1016/S0165-0270(99)00124-7

Ramadan, W., Eschenko, O., & Sara, S. J. (2009). Hippocampal Sharp Wave/Ripples during Sleep for Consolidation of Associative Memory. *PLoS ONE*, *4*(8), e6697. https://doi.org/10.1371/journal.pone.0006697

Ranck Jr, J. B. (1973). Studies on single neurons in dorsal hippocampal formation and septum in unrestrained rats: Part I. Behavioral correlates and firing repertoires. *Experimental neurology*, *41*(2), 462-531. https://doi.org/10.1016/0014-4886(73)90290-2

Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nature Reviews Neuroscience*, *7*(12), 967–975. https://doi.org/10.1038/nrn2022

Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews. Neuroscience*, *17*(3), 147–159. https://doi.org/10.1038/nrn.2015.30

Redondo, R. L., & Morris, R. G. M. (2011). Making memories last: the synaptic tagging and capture hypothesis. *Nature Reviews Neuroscience*, *12*(1), 17–30. https://doi.org/10.1038/nrn2963

Rich, P. D., Liaw, H.-P., & Lee, A. K. (2014). Large environments reveal the statistical structure governing hippocampal representations. *Science*, *345*(6198), 814–817. https://doi.org/ 10.1126/science.1255635

Robins, A. (1995). Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, *7*(2), 123-146. https://doi.org/10.1080/09540099550039318

Rodriguez, P. F., Aron, A. R., & Poldrack, R. A. (2006). Ventral–striatal/nucleus–accumbens sensitivity to prediction errors during classification learning. *Human Brain Mapping*, *27*(4), 306–313. https://doi.org/10.1002/hbm.20186

Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T. P., & Wayne, G. (2018). *Experience Replay for Continual Learning*. http://arxiv.org/abs/1811.11682

Roscow, E. L., Jones, M. W. & Lepora, N. F. (2019). Behavioural and computational evidence for memory consolidation based by reward-prediction errors. *BioRxiv.* https://doi.org/10.1101/716290

Rothschild, G. (2019). The transformation of multi-sensory experiences into memories during sleep. *Neurobiology of Learning and Memory*, *160*, 58–66. https://doi.org/10.1016/J.NLM.2018.03.019

Rothschild, G., Eban, E., & Frank, L. M. (2017). A cortical-hippocampal-cortical loop of information processing during memory consolidation. *Nature Neuroscience*, *20*(2), 251–259. https://doi.org/10.1038/nn.4457

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLOS Computational Biology*, *13*(9), e1005768. https://doi.org/10.1371/journal.pcbi.1005768

Sadacca, B. F., Jones, J. L., & Schoenbaum, G. (2016). Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *ELife*, *5*. https://doi.org/10.7554/eLife.13665

Sales, A. C., Friston, K. J., Jones, M. W., Pickering, A. E., & Moran, R. J. (2019). Locus Coeruleus tracking of prediction errors optimises cognitive flexibility: An Active Inference model. *PLOS Computational Biology*, *15*(1), e1006267. https://doi.org/10.1371/journal.pcbi.1006267

Salinas, A. G., Davis, M. I., Lovinger, D. M., & Mateo, Y. (2016). Dopamine dynamics and cocaine sensitivity differ between striosome and matrix compartments of the striatum. *Neuropharmacology*, *108*, 275-283. https://doi.org/10.1016/j.neuropharm.2016.03.049

Samanta, A., Alonso, A., & Genzel, L. (2020). Memory reactivations and consolidation: considering neuromodulators across wake and sleep. *Current Opinion in Physiology*, *15*, 120-127. https://doi.org/10.1016/j.cophys.2020.01.003

Sato, N., & Yamaguchi, Y. (2003). Memory encoding by theta phase precession in the hippocampal network. *Neural Computation*, *15*(10), 2379-2397. https://doi.org/10.1162/089976603322362400

Save, E., Nerad, L., & Poucet, B. (2000). Contribution of multiple sensory information to place field stability in hippocampal place cells. *Hippocampus*, *10*(1), 64–76. https://doi.org/10.1002/(SICI)1098-1063(2000)10:1<64::AID-HIPO7>3.0.CO;2-Y

Scarpetta, S., & de Candia, A. (2013). Neural Avalanches at the Critical Point between Replay and Non-Replay of Spatiotemporal Patterns. *PLoS ONE*, *8*(6), e64162. https://doi.org/10.1371/journal.pone.0064162

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, *26*(1), 3–8. https://doi.org/10.1002/hipo.22523

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). *Prioritized Experience Replay*. http://arxiv.org/abs/1511.05952

Schmidt, B., Papale, A., Redish, A. D., & Markus, E. J. (2013). Conflict between place and response navigation strategies: effects on vicarious trial and error (VTE) behaviors. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *20*(3), 130–138. https://doi.org/10.1101/lm.028753.112

Schonberg, T., Daw, N. D., Joel, D., O'Doherty, J. P., & Rangel, A. (2007). Reinforcement Learning Signals in the Human Striatum Distinguish Learners from Nonlearners during Reward-Based Decision Making. *Journal of Neuroscience*, *27*(47), 12860–12867. https://doi.org/10.1523/JNEUROSCI.2496-07.2007

Schouten, D. I., Pereira, S. I. R., Tops, M., & Louzada, F. M. (2017). State of the art on targeted memory reactivation: Sleep your way to enhanced cognition. *Sleep Medicine Reviews*, *32*, 123–131. https://doi.org/10.1016/J.SMRV.2016.04.002

Schultz, W, Dayan, P., Montague, P. R., Deichmann, R., Friston, K., & Dolan, R. J. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schultz, Wolfram. (1998). Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology*, *80*(1), 1–27. https://doi.org/10.1152/jn.1998.80.1.1

Schultz, Wolfram. (2016). Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience*, *18*(1), 23–32.

Schwartz, J. C., Diaz, J., Bordet, R., Griffon, N., Perachon, S., Pilon, C., ... & Sokoloff, P. (1998). Functional implications of multiple dopamine receptor subtypes: the D1/D3 receptor coexistence. *Brain research reviews*, *26*(2-3), 236-242. https://doi.org/10.1016/S0165-0173(97)00046-5

Schwindel, C. D., & McNaughton, B. L. (2011). Hippocampal–cortical interactions and the dynamics of memory trace reactivation. *Progress in Brain Research*, *193*, 163–177. https://doi.org/10.1016/B978-0-444-53839-0.00011-9

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*(1), 11–21. https://doi.org/10.1136/jnnp.20.1.11

Scudder, S. L., Baimel, C., Macdonald, E. E., & Carter, A. G. (2018). Hippocampal-evoked feedforward inhibition in the nucleus accumbens. Journal of Neuroscience, 38(42), 9091-9104. https://doi.org/10.1523/JNEUROSCI.1971-18.2018

Shen, B., & McNaughton, B. L. (1996). Modeling the spontaneous reactivation of experience-specific hippocampal cell assembles during sleep. *Hippocampus*, *6*(6), 685–692. https://doi.org/10.1002/(SICI)1098-1063(1996)6:6<685::AID-HIPO11>3.0.CO;2-X

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function. *Neuron*, *79*(2), 217–240. https://doi.org/10.1016/J.NEURON.2013.07.007

Shimp, C. P. (1970). A within-session effect after prolonged training in probability learning by rats. *Psychonomic Science*, *18*(3), 152–153. https://doi.org/10.3758/BF03332349

Shin, H., Lee, J. K., Kim, J., & Kim, J. (2017). Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems* (pp. 2990-2999). https://arxiv.org/abs/1705.08690

Siapas, A. G., & Wilson, M. A. (1998). Coordinated Interactions between Hippocampal Ripples and Cortical Spindles during Slow-Wave Sleep. *Neuron*, *21*(5), 1123–1128. https://doi.org/10.1016/S0896-6273(00)80629-7

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., … Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Singer, A. C., & Frank, L. M. (2009). Rewarded Outcomes Enhance Reactivation of Experience in the Hippocampus. *Neuron*, *64*(6), 910–921. https://doi.org/10.1016/j.neuron.2009.11.016

Skaggs, W. E., & McNaughton, B. L. (1996). Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science*, *271*(5257), 1870–1873. https://doi.org/ 10.1126/science.271.5257.1870

Skelin, I., Kilianski, S., & McNaughton, B. L. (2019). Hippocampal coupling with cortical and subcortical structures in the context of memory consolidation. *Neurobiology of Learning and Memory*, *160*, 21–31. https://doi.org/10.1016/J.NLM.2018.04.004

Soares-Cunha, C., Coimbra, B., David-Pereira, A., Borges, S., Pinto, L., Costa, P., ... & Rodrigues, A. J. (2016). Activation of D2 dopamine receptor-expressing neurons in the nucleus accumbens increases motivation. *Nature communications*, *7*(1), 1-11. https://doi.org/10.1038/ncomms11829

Sosa, M., Joo, H. R., & Frank, L. M. (2019). Dorsal and ventral hippocampus engage opposing networks in the nucleus accumbens. *BioRxiv*, 604116. https://doi.org/10.1101/604116

Spellman, T., Rigotti, M., Ahmari, S. E., Fusi, S., Gogos, J. A., & Gordon, J. A. (2015). Hippocampal-prefrontal input supports spatial encoding in working memory. *Nature*, *522*(7556), 309–314. https://doi.org/10.1038/nature14445

Spruston, N., & McBain, C. J. (2006). Structural and Functional Properties of Hippocampal Neurons. In *The Hippocampus Book* (pp. 133–202). https://doi.org/10.1093/acprof:oso/9780195100273.003.0005

Staddon, J. E. R., & Cerutti, D. T. (2003). Operant Conditioning. *Annual Review of Psychology*, *54*(1), 115–144. https://doi.org/10.1146/annurev.psych.54.101601.145124

Staresina, B. P., Alink, A., Kriegeskorte, N., & Henson, R. N. (2013). Awake reactivation predicts memory in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(52), 21159–21164. https://doi.org/10.1073/pnas.1311989110

Sterling, P., & Laughlin, S. (2015). *Principles of Neural Design*. Cambridge: MIT Press.

Stickgold, R. (2005). Sleep-dependent memory consolidation. *Nature*, *437*(7063), 1272–1278. https://doi.org/10.1038/nature04286

Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: evolving generalization through selective processing. *Nature Neuroscience*, *16*(2), 139–145. https://doi.org/10.1038/nn.3303

Stopper, C. M., & Floresco, S. B. (2011). Contributions of the nucleus accumbens and its subregions to different aspects of risk-based decision making. *Cognitive, Affective, & Behavioral Neuroscience*, *11*(1), 97–112. https://doi.org/10.3758/s13415-010-0015-9

Stott, J. J., & Redish, A. D. (2014). A functional difference in information processing between orbitofrontal cortex and ventral striatum during decision-making behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130472. https://doi.org/10.1098/rstb.2013.0472

Stuber, G. D., Hnasko, T. S., Britt, J. P., Edwards, R. H., & Bonci, A. (2010). Dopaminergic terminals in the nucleus accumbens but not the dorsal striatum corelease glutamate. *Journal of Neuroscience*, *30*(24), 8229-8233. https://doi.org/10.1523/JNEUROSCI.1754-10.2010

Studte, S., Bridger, E., & Mecklinger, A. (2017). Sleep spindles during a nap correlate with post sleep memory performance for highly rewarded word-pairs. *Brain and Language*, *167*, 28–35. https://doi.org/10.1016/J.BANDL.2016.03.003

Surmeier, D. J., Song, W. J., & Yan, Z. (1996). Coordinated expression of dopamine receptors in neostriatal medium spiny neurons. *Journal of neuroscience*, *16*(20), 6579-6591. https://doi.org/10.1523/JNEUROSCI.16-20-06579.1996

Sutherland, R. J., Weisend, M. P., Mumby, D., Astur, R. S., Hanlon, F. M., Koerner, A., … Hoesing, J. M. (2001). Retrograde amnesia after hippocampal damage: Recent vs. remote memories in two tasks. *Hippocampus*, *11*(1), 27–42. https://doi.org/10.1002/1098-1063(2001)11:1<27::AID-HIPO1017>3.0.CO;2-4

Sutton, R. S., & Barto, A. G. (1998). *Introduction to Reinforcement Learning*. Vol. 135. Cambridge: MIT Press.

Takahashi, Y. K., Langdon, A. J., Niv, Y., & Schoenbaum, G. (2016). Temporal Specificity of Reward Prediction Errors Signaled by Putative Dopamine Neurons in Rat VTA Depends on Ventral Striatum. *Neuron*, *91*(1), 182–193. https://doi.org/10.1016/J.NEURON.2016.05.015

Tecuapetla, F., Patel, J. C., Xenias, H., English, D., Tadros, I., Shah, F., ... & Koos, T. (2010). Glutamatergic signaling by mesolimbic dopamine neurons in the nucleus accumbens. *Journal of Neuroscience*, *30*(20), 7105-7110. https://doi.org/10.1523/JNEUROSCI.0265-10.2010

Tessler, C., Givony, S., Zahavy, T., Mankowitz, D. J., & Mannor, S. (2016). *A Deep Hierarchical Approach to Lifelong Learning in Minecraft*. http://arxiv.org/abs/1604.07255

Tang, L., Shafer, A. T., & Ofen, N. (2018). Prefrontal Cortex Contributions to the Development of Memory Formation. *Cerebral Cortex*, *28*(9), 3295–3308. https://doi.org/10.1093/cercor/bhx200

Thomas, M. S. C., & McClelland, J. L. (2008). Connectionist Models of Cognition. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 23–58). https://doi.org/10.1017/CBO9780511816772.005

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), i–109. https://doi.org/10.1037/h0092987

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Trouche, S., Koren, V., Doig, N. M., Ellender, T. J., El-Gaby, M., Lopes-dos-Santos, V., … Dupret, D. (2019). A Hippocampus-Accumbens Tripartite Neuronal Motif Guides Appetitive Memory in Space. *Cell*, *176*(6), 1393-1406.e16. https://doi.org/10.1016/J.CELL.2018.12.037

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, *26*(1), 1–12. https://doi.org/10.1037/h0080017

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G. (1999). The role of locus coeruleus in the regulation of cognitive performance. *Science*, *283*(5401), 549–554. https://doi.org/10.1126/science.283.5401.549

Valdés, J. L., McNaughton, B. L., & Fellous, J.-M. (2015). Offline reactivation of experience-dependent neuronal firing patterns in the rat ventral tegmental area. *Journal of Neurophysiology*, *114*(2), 1183–1195. https://doi.org/10.1152/jn.00758.2014

van de Ven, G. M., Trouche, S., McNamara, C. G., Allen, K., & Dupret, D. (2016). Hippocampal Offline Reactivation Consolidates Recently Formed Cell Assembly Patterns during Sharp Wave-Ripples. *Neuron*, *92*(5), 968–974. https://doi.org/10.1016/J.NEURON.2016.10.020

van der Meer, M. A. A., Johnson, A., Schmitzer-Torbert, N. C., & Redish, A. D. (2010). Triple Dissociation of Information Processing in Dorsal Striatum, Ventral Striatum, and Hippocampus on a Learned Spatial Decision Task. *Neuron*, *67*(1), 25–32. https://doi.org/10.1016/J.NEURON.2010.06.023

van der Meer, M. A. A., & Redish, A. D. (2009). Covert expectation-of-reward in rat ventral striatum at decision points. *Frontiers in Integrative Neuroscience*, *3*, 1. https://doi.org/10.3389/neuro.07.001.2009

van der Meer, M. A. A., & Redish, A. D. (2010). Expectancies in decision making, reinforcement learning, and ventral striatum. *Frontiers in Neuroscience*, *3*, 6. https://doi.org/10.3389/neuro.01.006.2010

van der Meer, M. A. A., & Redish, A. D. (2011). Theta phase precession in rat ventral striatum links place and reward information. *Journal of Neuroscience*, *31*(8), 2843–2854. https://doi.org/10.1523/JNEUROSCI.4869-10.2011

van Seijen, H., & Sutton, R. S. (2013). *Planning by Prioritized Sweeping with Small Backups*. http://arxiv.org/abs/1301.2343

Vanseijen, H., & Sutton, R. (2015, June). A deeper look at planning as learning from replay. In *International conference on machine learning* (pp. 2314-2322).

Vendrell-Llopis, N., Koralek, A., Costa, R., & Carmena, J. (2019). Ventral striatum uses a temporal difference rule for prediction during neuroprosthetic control. *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, 562–565. https://doi.org/10.1109/NER.2019.8716982

Villette, V., Malvache, A., Tressard, T., Dupuy, N., & Cossart, R. (2015). Internally Recurring Hippocampal Sequences as a Population Template of Spatiotemporal Information. *Neuron*, *88*(2), 357–366. https://doi.org/10.1016/J.NEURON.2015.09.052

Voorn, P., Vanderschuren, L. J. M. J., Groenewegen, H. J., Robbins, T. W., & Pennartz, C. M. A. (2004). Putting a spin on the dorsal–ventral divide of the striatum. *Trends in Neurosciences*, *27*(8), 468–474. https://doi.org/10.1016/J.TINS.2004.06.006

Vulkan, N. (2000). An Economist's Perspective on Probability Matching. *Journal of Economic Surveys*, *14*(1), 101–118. https://doi.org/10.1111/1467-6419.00106

Vyazovskiy, V. V., Olcese, U., Lazimy, Y. M., Faraguna, U., Esser, S. K., Williams, J. C., ... & Tononi, G. (2009). Cortical firing and sleep homeostasis. *Neuron*, *63*(6), 865-878. https://doi.org/10.1016/j.neuron.2009.08.024

Walsh, M. M., & Anderson, J. R. (2011). Learning from delayed feedback: neural responses in temporal credit assignment. *Cognitive, Affective, & Behavioral Neuroscience*, *11*(2), 131–143. https://doi.org/10.3758/s13415-011-0027-0

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., & de Freitas, N. (2016). *Sample Efficient Actor-Critic with Experience Replay*. http://arxiv.org/abs/1611.01224

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., … Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868. https://doi.org/10.1038/s41593-018-0147-8

Watkins, C. J. C. H. (1989). Learning form delayed rewards. *Ph. D. Thesis, King's College, University of Cambridge*. https://ci.nii.ac.jp/naid/10007782517/

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3–4), 279–292. https://doi.org/10.1007/BF00992698

Weinstock, S., North, A., Brody, A. L., & Loguidice, J. (1965). Probability learning in the T maze with noncorrection. *J ournal of comparative and physiological psychology, 60(1),* 76-81. https://doi.org/10.1037/h0022368

Wikenheiser, A. M., & Redish, A. D. (2015). Hippocampal theta sequences reflect current goals. *Nature Neuroscience*, *18*(2), 289–294. https://doi.org/10.1038/nn.3909

Wikenheiser, A. M., & Redish, A. D. (2013). The balance of forward and backward hippocampal sequences shifts across behavioral states. Hippocampus, 23(1), 22-29. https://doi.org/10.1002/hipo.22049

Wilkinson, L., Tai, Y. F., Lin, C. S., Lagnado, D. A., Brooks, D. J., Piccini, P., & Jahanshahi, M. (2014). Probabilistic classification learning with corrective feedback is associated with in vivo striatal dopamine release in the ventral striatum, while learning without feedback is not. *Human Brain Mapping*, *35*(10), 5106–5115. https://doi.org/10.1002/hbm.22536

Wilson, M. A., & McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Science*, *261*(5124), 1055–1058. https://doi.org/10.1126/science.2911737

Worthy, D. A., & Todd Maddox, W. (2014). A comparison model of reinforcement-learning and win-stay-lose-shift decision-making processes: A tribute to W.K. Estes. *Journal of Mathematical Psychology*, *59*, 41–49. https://doi.org/10.1016/J.JMP.2013.10.001

Wu, C.-T., Haggerty, D., Kemere, C., & Ji, D. (2017). Hippocampal awake replay in fear memory retrieval. *Nature Neuroscience*, *20*(4), 571–580. https://doi.org/10.1038/nn.4507

Yang, C. R., & Mogenson, G. J. (1984). Electrophysiological responses of neurones in the nucleus accumbens to hippocampal stimulation and the attenuation of the excitatory responses by the mesolimbic dopaminergic system. *Brain research*, *324*(1), 69-84. https://doi.org/10.1016/0006-8993(84)90623-1

Yorgason, J. T., Zeppenfeld, D. M., & Williams, J. T. (2017). Cholinergic interneurons underlie spontaneous dopamine release in nucleus accumbens. *Journal of Neuroscience*, *37*(8), 2086-2096. https://doi.org/10.1523/JNEUROSCI.3064-16.2017

Yu, J. Y., Kay, K., Liu, D. F., Grossrubatscher, I., Loback, A., Sosa, M., … Frank, L. M. (2017). Distinct hippocampal-cortical memory representations for experiences associated with movement versus immobility. *ELife*, *6*. https://doi.org/10.7554/eLife.27621

Záborszky, L., Alheid, G. F., Beinfeld, M. C., Eiden, L. E., Heimer, L., & Palkovits, M. (1985). Cholecystokinin innervation of the ventral striatum: a morphological and radioimmunological study. *Neuroscience*, *14*(2), 427-453. https://doi.org/10.1016/0306-4522(85)90302-1

Zahm, D. S. (2000). An integrative neuroanatomical perspective on some subcortical substrates of adaptive responding with emphasis on the nucleus accumbens. *Neuroscience & Biobehavioral Reviews*, *24*(1), 85-105. https://doi.org/10.1016/S0149-7634(99)00065-2

Zheng, C., Bieri, K. W., Trettel, S. G., & Colgin, L. L. (2015). The relationship between gamma frequency and running speed differs for slow and fast gamma rhythms in freely behaving rats. *Hippocampus*, *25*(8), 924–938. https://doi.org/10.1002/hipo.22415

Ziv, Y., Burns, L. D., Cocker, E. D., Hamel, E. O., Ghosh, K. K., Kitch, L. J., … Schnitzer, M. J. (2013). Long-term dynamics of CA1 hippocampal place codes. *Nature Neuroscience*, *16*(3), 264–266. https://doi.org/10.1038/nn.3329