Model-Free and Model-Based Influence on Choice in Rodents and Interactions between Hippocampus and Dorsomedial Prefrontal Cortex during Deliberation

A Dissertation
Submitted to the Faculty of the
University of Minnesota
by

Brendan Hasz

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Advisor: A. David Redish

January 2020

Acknowledgements

Most importantly, I'd like to thank Dr. A. David Redish for taking me on as a graduate student, and for providing the mentorship, the environment, and the freedom needed to complete the research described in this thesis, and to learn what I wanted to learn, while still keeping me on track to graduate within the decade.

I'd also like to thank the members of my thesis committee Dr. Matt Chafee, Dr. Tay Netoff, and Dr. Matt Johnson, for being friendly and supportive while providing suggestions and advice.

A huge thanks to Ayaka Sheehan, Chris Boldt, and Kelsey Seeland for building all the tetrode and probe drives, general technical support, answering all my questions with infinite patience, providing impromptu medical assistance, for the tea, and just for generally being awesome and fun coworkers. Thanks again to Ayaka for doing all the histology and making sure my electrodes were (or weren't!) all in the right place.

Thanks to Patrick Crowe, Charlie Jackson, Onni Rauhala, and Daniel Min for help running rats.

Also thanks to Yannick Breton for writing the code to track rat head position from video, and to other previous lab members who contributed to parts of the Redish lab code set which I used.

A big thanks to Redish lab members past and present whom I haven't already mentioned: Nate, Geoff S., Paul, Brandy, Evan, Cody, Brian, Geoff D., Olivia, and everyone else! You've all made this a fun place to be. And thanks to all my friends in the GPN for dragging me out of lab when needed!

Thanks to the IGERT Neuroengineering program not only for financial support, but for the weekly meetings with a fun group of people.

And, of course, thanks to the rats. Even that one who tried to bite me.

The work in this dissertation was funded by NSF IGERT Neuroengineering grant DGE-1069104, NIH R01-MH112688, and NIH R01-MH080318.

Abstract

Decision making is driven by multiple, somewhat independent systems within the brain. One of these systems makes slow, deliberative decisions, and is thought to be driven by a model-based neural algorithm, in that it learns an internal model of the world which it uses to make decisions. Another system makes fast, habitual choices, and is hypothesized to depend on a model-free neural algorithm, in that it does not learn a model of the world, but simply stores state-action-reward associations. While the habitual system is relatively well-studied, the neural underpinnings of the deliberative system are less clear. Specifically, it is not known how areas comprising the deliberative system, such as prefrontal cortex and the hippocampus, share information on fast timescales. Also, representations of contingency information in prefrontal areas have previously been impossible to disambiguate from the encoding of other time-varying information. In this thesis, we adapted for rats a task which enabled the dissociation of model-based from model-free influence on choice, and we found evidence for both model-based and model-free control. We also developed a simpler task which caused rats to repeatedly transition between deliberative and habitual modes. On this second task, we found that both dmPFC and CA1 encoded information about task contingencies, while simultaneously representing unrelated time-varying information. Lastly, we examined interactions between dmPFC and CA1 on fast timescales, and found that both areas represented prospective information simultaneously, but that the content of this prospective information was not always identical between the two areas. Activity in dmPFC predicted whether HPC would represent prospective information on broad timescales, and prospective representation in HPC changed reward encoding in dmPFC on faster, sub-second timescales. Our work begins to bridge the neural underpinnings of decision making in rodents and the algorithms by which they select actions, confirms that the deliberative system represents contingency information, and uncovers asymmetries in the transfer of information between dmPFC and HPC.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

How does the brain compute which actions will realize its goals? Understanding the neural systems which perform decision-making is critical not just for better understanding the brain, but also for developing treatments and therapies for a wide range of psychopathologies which arise due to problems in decision making, including addiction, obsessive-compulsive disorder, and attention deficit-hyperactivity disorder.

Current theories suggest that there are multiple, somewhat independent subsystems within the brain that contribute to decision-making (O'Keefe and Nadel, 1978a; Adams and Dickinson, 1981; Sloman, 1996; Dayan and Balleine, 2002; Lieberman, 2003; Loewenstein and O'Donoghue, 2004; Balleine et al., 2008; van der Meer et al., 2012; Kahneman, 2011; Redish, 2013; Dolan and Dayan, 2013). These systems are thought to use different algorithms to select actions, and may be used at different times, or even be active simultaneously and give rise to conflicting decisions. Some research dissociates deliberative or goal-directed decision-making from habitual decision-making systems (van der Meer et al., 2012; Redish, 2013), work in humans separately identifies "model-based" and "model-free" influences on decision-making (Gläscher et al., 2010; Daw et al., 2011), other research distinguishes place from response strategies (O'Keefe and Nadel, 1978a; Packard and McGaugh, 1996; Redish, 1999; Yin and Knowlton, 2004), other work separates associative from rule-based systems (Sloman, 1996), others distinguish "fast" and "slow" systems (Kahneman, 2011), and still others identify reflexive from reflective processes (Lieberman, 2003). A central theme to these dissociations is separating a fast, automatic, habit-like system from a slow, intentional, deliberative system. Although there are likely more than just two such systems (Redish, 2013), this thesis will focus on the deliberative and habitual systems, which are thought to rely on "model-based" and "model-free" neural algorithms, respectively.

The deliberative system is thought to employ a model-based algorithm to make decisions. Model-based algorithms learn and store an internal model of the world, and use this model to make more intelligent decisions (Doll et al., 2012). Model-based algorithms tend to perform better in environments where either information is limited, or when the environment is changing quickly (Gläscher et al., 2010; Daw et al., 2011). However, because the use of the model-based algorithm relies on repeated simulations of the internal model, decisions made with this system are comparatively slow, which can be an important drawback in situations or environments where speed is required (Keramati et al., 2011).

On the other hand, the neural system giving rise to habitual or procedural behaviors is thought to be supported by a model-free neural algorithm. Unlike model-based algorithms, model-free algorithms do not learn a model of the world to store state transition probabilities (thus the name!), and instead simply store associations between states, actions, and the resulting rewards. Essentially, these algorithms store the expected value of taking any action in any given state, and update those expected reward values as the agent performs those actions and experiences some amount of reward as a result (Watkins, 1989; Rummery and Niranjan, 1994; Sutton and Barto, 1998). Because this class of algorithms is association-based, instead of simulation-based as with model-based algorithms, the computations required to make decisions are very fast (Keramati et al., 2011). However, the drawback of this faster style of algorithm is that they are relatively inflexible compared to model-based algorithms. The state-action-reward associations in model-free systems are only able to be updated as a result of direct experience, and so agents must repeatedly experience sequences of events. Model-based algorithms, on the other hand, are able to dynamically update valuations because they are able to synthesize knowledge of parts of an environment to make more intelligent decisions, without having to directly experience the entire chain of events sequentially (Doll et al., 2012).

However, most work dissociating model-based from model-free influence on decision making has been performed in humans, where measurements of information representation on fast timescales is difficult. In rodents, where simultaneously recorded ensembles of single units make possible the measurement of information representation on fast time scales, research has focused on separating deliberative neural systems from those neural systems giving rise to habitual or procedural behavior. It is thought that a model-based neural algorithm underlies the deliberative system, and a model-free neural algorithm drives habitual behavior, but the accuracy of this hypothesized parallel remains unclear.

The neural correlates of the habitual system have been well-studied both experimentally (Packard and McGaugh, 1996; Schultz et al., 1997; Jog et al., 1999; Schmitzer-Torbert and Redish, 2004; Yin and Knowlton, 2004) and theoretically (Niv et al., 2006; Frank, 2011), and are thought to involve dorsal striatal areas, along with the rest of the basal ganglia, the thalamus, and motor and sensory cortices. In contrast, the neural underpinnings of the deliberative system appear to be more complex and are less well-understood (Doll et al., 2012; van der Meer et al., 2012). In rodents, some work observes the representation of prospective information by ensembles in the hippocampus (HPC) during deliberation at choice points (Johnson and Redish, 2007), which is thought to correspond to the simulation of the outcomes of candidate actions using an internal model stored in part by the hippocampus. Furthermore, neural activity thought to corresponding to the estimation of the value of these simulated outcomes has been observed in ventral striatum (van der Meer and Redish, 2010) and in orbitofrontal cortex (Rich and Wallis, 2016; Wallis, 2018). However, it's unclear how these valuations are then used to select between candidate actions, and also where and how internal simulations of candidate action outcomes are initiated. If model-based theories for the operation of the deliberative system are correct, then presumably some brain areas are responsible for detecting the need for deliberative control, instigating the internal simulation of outcomes associated with candidate actions, storing in working memory the estimated value of those simulated outcomes, and after the value of multiple candidate actions has been estimated, using that value information stored in working memory to make a decision as to which action to take.

Candidate brain regions for performing some or all of these roles include the various subregions of the prefrontal cortex (PFC). The hippocampus (HPC) and PFC, along with other structures, are thought to form an information-processing loop where top-down contextual signals from PFC influence encoding in HPC, and information retreival by HPC informs representations in PFC. This loop may also be responsible for the initiation, simulation, and evaluation of candidate actions (van der Meer et al., 2012). Various regions of the prefrontal cortex (PFC) have long been thought to mediate executive function (Miller and Cohen, 2001; Dalley et al., 2004; Kesner and Churchwell, 2011). The PFC, specifically the anterior cingulate cortex, has been implicated in conflict detection (Haddon and Killcross, 2005, 2006; Marquis et al., 2007; Dwyer et al., 2010), suggesting it may be responsible for detecting the need for deliberative control. Also, the PFC plays an active role in the storage and recall of working memories (Tronel and Sara, 2003; Ragozzino and Kesner, 1998; Delatour and Gisquest-Verrier, 1999; Cowen and McNaughton, 2007; Yoon et al., 2008; Horst and Laubach, 2009; Euston et al., 2012; Preston and Eichenbaum, 2013; Urban et al., 2014), which may also translate to the storage of internally simulated outcome valuations. It is theorized that PFC may initiate the internal construction of hypothetical situations (Hassabis and Maguire, 2009; van der Meer et al., 2012; Wang et al., 2015).

Specifically, the dorsomedial prefrontal cortex (dmPFC) also represents information about environmental contingencies or behavioral strategies (Balleine and Dickinson, 1998; Jung et al., 1998; Wallis et al., 2001; Ragozzino et al., 2003; Floresco et al., 2008; Young and Shapiro, 2009; Hyman et al., 2012; Mante et al., 2013; Powell and Redish, 2014; Ma et al., 2016). This contingency information, or hidden information which must be learned through experience and stored in working memory and used to make optimal decisions, is very similar in spirit to the abstract information about the world thought to be required for implementing a model-based algorithm.

However, it is hard to parse out how much of this apparent contingency representation is due to actual contingency representation, as opposed to an artifact of representational drift over time. Most work examining latent contingency representations in prefrontal areas employ tasks where the contingencies are present in blocks of trials or time. This is because if the contingencies were to be cued, it would be completely impossible to distinguish representations of contingency from representations of sensory information. Unfortunately, in removing one confound, this block-like task structure gives rise to a second confound: time.

Neural activity in both dorsomedial prefrontal cortex and hippocampus are known to change slowly over time (Mankin et al., 2012; Hyman et al., 2012; Ziv et al., 2013).

If then, neural activity is changing over time, then it is difficult to say whether differences between neural activity across task blocks is due to encoding of information specific to those blocks (of interest here, contingency information), or whether those differences are due simply to some unrelated random drift over time. Studies attempting to disambiguate these two contributions to neural activity use decoding, ensemble correlation, or clustering approaches to determine whether ensemble activity represents contingency information, for example Malagon-Vina et al. (2018). However, these decoding approaches suffer from the aforementioned inability to disambiguate contingency representation (when contingencies are presented in blocks of time) from unrelated representational drift over time. Alternative approaches have been taken which find sudden representational shifts coincident with contingency changes (Rich and Shapiro, 2009; Durstewitz et al., 2010; Karlsson et al., 2012; Powell and Redish, 2016).

Assuming these brain areas are indeed representing abstract contingency information, it is then also unknown how the timing of representational changes differ between brain areas representing this information. For the case of dorsomedial prefrontal cortex and the hippocampus, theories of a spatial working memory loop between the two areas certainly suggest that contingency information would appear first in dorsomedial prefrontal cortex, and then make its way into hippocampal representations, but it is unclear how quickly this transfer of information occurs. Also, the presence of unrelated representational drift over time further complicates any measurement of timing differences, as it is unknown how the drift rates differ between hippocampus and prefrontal cortex.

It is also less well studied how HPC and dmPFC share other types of information (e.g. about reward, candidate actions, and location) on fast timescales. Theories of the deliberative system indicate that prefrontal areas detect a need for deliberative control, instigate internal simulations of the outcomes of candidate actions, keep track of the valuations of those outcomes, and use that information to decide which candidate action to enact. Therefore, likely candidates for information being passed between dmPFC and HPC include information about candidate actions, location, and reward. Previous work has discovered that hippocampal ensembles represent non-local spatial information which appears to correspond to internal simulations of candidate actions (Johnson and Redish, 2007), and other areas such as ventral striatum and orbitofrontal cortex represent value in ways suggesting they may be estimating the value of these internally simulated outcomes (van der Meer and Redish, 2010; Rich and Wallis, 2016). But what instigates these internal simulations? If dmPFC plays this role, then it should be possible to predict from activity in dmPFC whether non-local information is about to be represented in HPC. Also, if dmPFC is keeping track of predicted value of candidate actions, then when the outcomes of these candidate actions are represented in HPC (which, in theory, cause value representations in other areas corresponding to the estimated subjective value of those simulated outcomes) should have an effect on reward encoding in dmPFC. However, work involving simultaneous ensemble recordings in both dmPFC and HPC have not yet investigated whether information encoding in these two areas occurs in this way on fast timescales.

In this thesis we first examine the model-free and model-based influences on rat behavior, and then further the representation of task-relevant information in dmPFC and CA1, and how they may contribute to the model-based deliberative system.

- Chapter 2 describes a task we adapted which is able to distinguish the contributions of model-based from model-free influences on rodent decisions in spatial mazes, and examines how rat decisions are explained by a combination of model-based and model-free influences, but finds that this spatial version of the task was not optimal for measuring trial-by-trial influences on rat choice, due mainly to its complexity.

- Chapter 3 introduces a different, simpler task we designed to study repeated transitions between deliberative and habitual decision-making modes, and examines how behavioral correlates of deliberation correspond to rats' uncertainty as to the task contingencies.

- Chapter 4 develops an analysis to disambiguate the contributions of contingency encoding from representational drift over time, demonstrates that both CA1 and dmPFC encode contingency information while simultaneously displaying representational drift over time, and examines the timing of these changes.

- Chapter 5 investigates the representation of spatial and reward information in dmPFC and CA1, and demonstrates that both areas represent prospective spatial information simultaneously, while not always representing identical locations, and that activity in dmPFC predicts non-local representation by HPC ensembles on broad timescales, while prospective activity in HPC effects reward representations in dmPFC on fast timescales.

- Chapter 6 summarizes our findings and their significance, and discusses potential avenues and challenges of future work in this area.

Chapter 2

Model-based and Model-free Decision Making on a Two-Step Task

The work discussed in this chapter has been previously reported in Hasz and Redish (2018).

2.1 Introduction

Current theories suggest that decision-making arises from multiple subsystems within the brain. Each system is thought to use different algorithms to select actions based on external, and sometimes internal, information. However, literatures using different experimental species have dissociated different types of decision-making systems in different ways (O'Keefe and Nadel, 1978a; Adams and Dickinson, 1981; Sloman, 1996; Dayan and Balleine, 2002; Lieberman, 2003; Loewenstein and O'Donoghue, 2004; Balleine et al., 2008; van der Meer et al., 2012; Redish, 2013; Dolan and Dayan, 2013).

Studies of rodent navigation through spatial mazes have revealed a dichotomy between deliberative behavior and procedural behavior. Deliberative behaviors are thought to arise from the use of some internal evaluation of the expected state of the world, or "cognitive map" (Muenzinger and Gentry, 1931; Tolman, 1939; O'Keefe and Nadel, 1978a). These deliberative behaviors are identified by the use of "place strategies", when rodents make decisions based on place or goal locations (Packard and McGaugh, 1996; Schmidt et al., 2013; Gardner et al., 2013; Redish, 2016). Deliberation is thought to involve an entire ensemble of brain areas, including the hippocampus, other more associative brain areas such as prefrontal cortex and orbitofrontal cortex, basal ganglia structures such as the ventral striatum, thalamic structures such as nucleus reuniens, and more (Redish, 1999; van der Meer et al., 2012).

In contrast, procedural behavior is a much faster process thought to be driven by habits. In rodents, procedural behavior is characterized by "response strategies", where animals make decisions based on relatively simple stimulus-action associations (Packard and McGaugh, 1996; Yin and Knowlton, 2004). Unlike deliberation, procedural behaviors are thought to be driven primarily by motor cortical and basal ganglia structures such as the dorsolateral striatum (Packard and McGaugh, 1996; Jog et al., 1999; Yin and Knowlton, 2004).

On the other hand, studies in humans dissociate decision-making behavior based on how subjects make choices consistent with those of "model-based" and "model-free" learning algorithms (Gläscher et al., 2010; Daw et al., 2011; Gillan et al., 2011; Wunderlich et al., 2012; Otto et al., 2013b,a; Eppinger et al., 2013; Skatova et al., 2013; Schad et al., 2014; Gillan et al., 2014; Sebold et al., 2014; Otto et al., 2015; Gillan et al., 2015; Voon et al., 2015; Deserno et al., 2015; Radenbach et al., 2015; Sharp et al., 2015; Doll et al., 2016; Decker et al., 2016), though some of this work has been done in rats on simplified tasks (Miller et al., 2013, 2014, 2017). However, some work suggests these simplified tasks used for rodents are unable to truly separate model-based from model-free influences on decisions (Akam et al., 2013).

Model-free algorithms were originally developed in the context of machine reinforcement learning. Some early versions of these algorithms include "Q-learning" (Watkins, 1989) and "SARSA" (Rummery and Niranjan, 1994). This class of algorithm learns the expected value of taking any given action in any given state. Essentially, the algorithm stores a lookup table of the expected reward associated with state-action pairs. Values in this table are updated according to the rewards the agent actually experiences, with the hope that over time they come to approximate the true values associated with each state-action pair. At inference time (when the agent needs to make a decision), the algorithm simply looks up the available actions for a given state and their estimated values, and chooses the action with the highest expected reward for that state. As this algorithm has constant time complexity (assuming the number of available actions is constant), it is very fast (Keramati et al., 2011).

However, model-free algorithms suffer from an important limitation: because they only update their reward expectations according to experiences the agent has already had, this class of algorithm is inflexible and performs poorly when contingencies change. That is, when state-action rewards remain relatively constant, but the relationships between states change, model-free algorithms must re-learn the expected rewards from "scratch". This is because model-free algorithms do not contain any sort of model of the world, but only associations between state-action pairs and reward.

Model-based algorithms, on the other hand, do store models of the world, and therefore are able to use this information to handle more dynamic environments. Early versions of these algorithms include the Dyna architecture (Sutton, 1991), and prioritized sweeping (Moore and Atkeson, 1993). Model-based algorithms are similar to model-free algorithms in that they too learn the expected reward associated with state-action pairs. However, they also build an internal model of the agent's environment which can be used to make more optimal decisions, especially in the face of dynamic contingencies. Specifically, these models are usually instantiations of a Markov decision process, and store the transition probabilities associated with each state-action pair. That is, they learn not just the amount of reward the agent can expect by performing a given action in a given state, but also how likely that action is to cause the state to change to any other given state. This information can then be used to evaluate on-line the tree of possible futures given different potential actions. This dynamic on-line evaluation of expected action outcomes allows an agent to more dynamically compute expected rewards, even when a given action (or chain of actions) hasn't yet been observed by the agent to lead to large rewards. While model-based algorithms allow for more flexibility and optimal learning, they are far more computationally demanding. Especially in the case of a large number of possible future states, the number of possible paths through those future states becomes vast, requiring much more computation at inference time (Keramati et al., 2011).

Recent behavioral and magnetic imaging work studying human subjects has investigated how human decisions and neural activity may be related to model-based and model-free algorithms. Much of this work employs a two-step task which is able to dissociate between decisions made by model-based algorithms from those made by model-free algorithms (Gläscher et al., 2010; Daw et al., 2011). Briefly, the task involves two sequential binary choices, where the first choice probabilistically controls which of two decisions will be presented for the second choice. Furthermore, the rewards associated with actions taken at the second choice change over time. This task is able to dissociate model-based from model-free decision making because model-based algorithms are able to use knowledge of the task structure and transition probabilities to update reward expectations at the first choice, while model-free algorithms are not. More details on the two-step task will be given in Section 2.2.

This body of work in humans has found evidence that human decisions on the two-step task are consistent with those of model-based and model-free algorithms (Gläscher et al., 2010). Furthermore, human brain areas such as the ventral striatum and prefrontal cortex appear to activate in ways consistent with reward prediction errors in model-based and model-free algorithms (Gläscher et al., 2010; Daw et al., 2011). Further work finds that various factors and disorders can disrupt the balance between model-free and model-based influences on human decision making. For example, obsessive-compulsive disorder appears to cause individuals to make decisions which are more consistent with a model-free strategy (Gillan et al., 2011, 2014; Voon et al., 2015). Also, subjects with alcohol dependence show weaker a influence of the model-based system (Sebold et al., 2014), while the acute effect of alcohol administration has been found to do the opposite (Obst et al., 2018). Individuals displaying higher levels of cognitive control or those with more working memory appear more model-based, and individuals are unable to behave as model-based when working memory is allocated elsewhere (Otto et al., 2013b,a; Schad et al., 2014; Otto et al., 2015). Dopamine appears to play an integral role in either the balance between the model-based and model-free systems, or the functioning of the model-based system. Increased dopamine corresponds to more model-based-like behavioral strategies, whether this increase in dopamine levels was experimentally increased via the administration of L-DOPA (Wunderlich et al., 2012; Sharp et al., 2015), or the amount of naturally-occuring dopamine as measured by F-DOPA positron emission tomography (Deserno et al., 2015) or genetic indicators (Doll et al., 2016). Age also has been found to play a role in determining the balance between model-based and model-free strategies (Eppinger et al., 2013; Decker et al., 2016), and using model-based strategies appears to defend against habit formation (Gillan et al., 2015). Stress leads to a decrease in the ability to make model-based choices (Radenbach et al., 2015), and some work has even found a relationship between an individual's extrovertedness and the balance between model-based and model-free influences on their choices (Skatova et al., 2013).

How do the model-free and model-based algorithms relate to procedural and deliberative behavior and neural activity? The procedural system is hypothesized to be driven by a model-free neural mechanism, in that it is not thought to actually use any internal model of the world to make decisions, but rather caches the expected best action for each given state.

Schultz et al. (1997) first provided evidence that neural activity in monkeys actually reflected internal variables of reinforcement learning models like the model-free algorithm (Sutton and Barto, 1998). Specifically, they found that dopaminergic cells in the ventral tegmental area (VTA) and the substantia nigra are tonically active, and did not change their baseline firing rates when monkeys received as much reward as they expected. However, when monkeys received unexpected rewards, there was a burst in the activity of the dopaminergic cells, and when an expected reward was omitted, there was a transient decrease in the firing rate of the dopaminergic units (Schultz et al., 1997). This behavior is consistent with the "delta signal" used in reinforcement learning algorithms (Sutton and Barto, 1998). This delta signal carries information about the difference between the algorithm's current estimate of the reward associated with a specific state-action pair, and the actual amount of reward that was received at that timepoint.

Dopamine release is known to affect plasticity of corticostriatal synapses, and in the context of habitual behavior (Calabresi et al., 2007), specifically the projections from sensory association and motor areas carrying state and motor plan information to dorsolateral striatal areas. The striatum, specifically the dorsolateral aspects of the striatum, are important for forming habits (Yin and Knowlton, 2004), and show bursts of activity at the initiation of habitual action chains (Jog et al., 1999). The striatum plays a key role in the basal ganglia circuit which gates action initiation and selection. Theories have suggested that changes in the amount of dopamine released onto corticostriatal synapses control the strength of these synapses, and therefore are able to tune how strongly an action is initiated (or silenced) upon input representing specific situations from sensory association areas (Swanson, 2000; Niv et al., 2006; Frank, 2011). This hypothesized system is very similar to the state-action pair reward associations of the model-free algorithm, and thus it is hypothesized that the neural system generating habitual behaviors implements a model-free algorithm, or something very similar.

The procedural system makes decisions quickly, but these decisions are habit-like and inflexible once learned. That is, for the procedural system, un-learning a decision-making policy requires a large amount of training relative to the deliberative system (Niv et al., 2006; Keramati et al., 2011; van der Meer et al., 2012). Although model-free algorithms are not necessarily slower to change their policies than model-based ones (this speed is primarily dependent on the learning rate), model-free algorithms do suffer from the limitation that the agent must observe the reward outcomes of a sequence of actions before updating their reward beliefs. In complex environments where the contingencies change, but not necessarily the rewards associated with taking actions in (potentially latent) states, model-free systems require many more experiences than do model-based systems to accurately update the estimated reward associated with state-action pairs. Thus, procedural decision making is thought to be driven by a model-free neural algorithm. Like such algorithms, procedural decisions do not quickly reflect changes in contingencies or state transition probabilities.

In contrast, the deliberative system is hypothesized to employ the model-based neural mechanism: it is thought that the deliberative system stores and evaluates an internal model of the world, based on contingencies or latent states, to estimate the outcomes of potential actions. The storage of this internal model has been proposed to reside in the hippocampus (Johnson and Redish, 2007; Redish, 2016) and perhaps also in sub-regions of the prefrontal cortex such as orbitofrontal cortex (Wikenheiser and Schoenbaum, 2016; Zhou et al., 2019). While the hippocampus has long been known to play a role in memory (Scoville and Milner, 1957), more recently it has been discovered that the hippocampus sometimes represents the potential outcomes of candidate actions while subjects deliberate (Johnson and Redish, 2007; Simon and Daw, 2011; Doll et al., 2015; Brown et al., 2016). The orbitofrontal cortex is also thought to represent information about the "cognitive map" (Wikenheiser and Schoenbaum, 2016; Zhou et al., 2019), though it is unclear what aspects of environment representation and simulation occur in orbitofrontal cortex and which occur in hippocampus, or how much of these roles are shared between the two structures. Prefrontal cortex is important for working memory (Ragozzino and Kesner, 1998; Delatour and Gisquest-Verrier, 1999; Cowen and McNaughton, 2007; Yoon et al., 2008; Horst and Laubach, 2009; Urban et al., 2014), and plays a role in decision-making and generating goal-directed (as opposed to habitual) actions (Seamans et al., 1995; Killcross and Coutureau, 2003; Matsumoto et al., 2003; Matsumoto and Tanaka, 2004; Hok et al., 2005; St. Onge and Floresco, 2009). Furthermore, ventral aspects of the striatum represent reward-related information while hippocampus represents potential outcomes of candidate actions (van der Meer and Redish, 2010), and so it is thought that during deliberation the ventral striatum plays the role of a "critic" to the hippocampus' "actor" (van der Meer et al., 2012). Taken together, this system has been hypothesized to perform internal simulations of a world model, and the evaluation of their simulated outcomes, to decide which actions to take. This storage of action outcomes in the context of environmental dynamics and hypothetical-based evaluation of action optimality is very similar to the model-based algorithm, and therefore it is believed that a model-based neural algorithm underlies deliberative decision-making and behaviors (Doll et al., 2012; Daw and Dayan, 2014; van der Meer et al., 2012; Redish, 2016).

This hypothesized internal model learns not only the expected reward for each state-action pair in the environment, but also learns the relationships between states - information the procedural (and putatively model-free) system does not represent or use. That information is thought to be integrated on-line in order to make more optimal decisions, even in completely new situations (Adams and Dickinson, 1981; van der Meer et al., 2012). However, like with the model-based algorithm, one key drawback of deliberation is that it requires more time and cognitive effort than the procedural system, because it requires both the repeated simulation and the evaluation of an internal model. It is hypothesized that the brain employs some sort of trade-off between fast, inflexible procedural strategies and slow, more flexible deliberative strategies (Keramati et al., 2011).

In addition to using the identification of place strategies, deliberation has also been identified in rodents by the presentation of a specific behavior termed "vicarious trial and error." Vicarious trial and error (VTE) is a behavior where rats pause at choice points of a maze, and look back and forth down each path as if deliberating over which path to take (Muenzinger and Gentry, 1931; Tolman, 1939; Redish, 2016). VTE behaviors are thought to occur during internal deliberative processes: the evaluation of an internal model of the world, which corresponds to a model-based neural mechanism (Redish, 2016).

During procedural behavior rats do not display VTE, and their paths through the choice points are instead highly regular and stereotyped (Packard and McGaugh, 1996; Jog et al., 1999; Schmitzer-Torbert and Redish, 2002; van der Meer et al., 2012; Smith and Graybiel, 2013; Schmidt et al., 2013). Again, the procedural system generating this stereotyped behavior is hypothesized to employ a model-free learning algorithm (O'Keefe and Nadel, 1978a; Jog et al., 1999; Yin and Knowlton, 2004; Frank, 2011; Redish, 2016). Animals usually display deliberative behavior early in training, which transitions to more stereotyped behavior with experience on a given task (Packard and McGaugh, 1996; Gardner et al., 2013; Schmidt et al., 2013; Redish, 2016).

While the model-based system is hypothesized to correspond to the deliberative system, and the model-free system to the procedural system, research has not actually mechanistically linked the hypothesized underlying algorithms to neural activity or behavior in rodents. The model-based/model-free dichotomy has been evaluated using tasks which differentiate decisions based on the apparent presence of knowledge about relations between states, information which only the model-based system stores (Daw et al., 2011; Doll et al., 2012). In contrast, the deliberative/procedural dichotomy has been evaluated using behavioral markers such as place/response strategies and VTE, but have not tied these behaviors to model-based or model-free choices. Both of these literatures have been very successful in dissociating two types of decision-making, but it is unknown how they correspond to each other.

Furthermore, the existence of multiple decision-making systems within the same agent raises the question of how an organism makes a single coherent action when multiple systems are contributing to a decision, potentially in conflicting ways. That is to say: how is an agent which consists of multiple decision-making systems, some of which make different decisions at the same time, able to come to one single decision which is eventually executed by the agent? Work in humans has assumed a subject-specific static weight between model-based and model-free influence (Gläscher et al., 2010; Daw et al., 2011; Gillan et al., 2011; Wunderlich et al., 2012; Otto et al., 2013b,a; Eppinger et al., 2013; Skatova et al., 2013; Schad et al., 2014; Gillan et al., 2014; Sebold et al., 2014; Otto et al., 2015; Gillan et al., 2015; Voon et al., 2015; Deserno et al., 2015; Radenbach et al., 2015; Sharp et al., 2015; Doll et al., 2016; Decker et al., 2016). For example, according to this hypothesis the model-free system contributes to all decisions with, say, 40% weight, and model-based with 60% weight. This may not be the case – anecdotal and introspective evidence would suggest that sometimes, one uses nearly entirely habitual control (say, when turning on a light switch when entering a familiar room), while at other times one uses nearly wholly deliberative control (say, when deciding which college to attend!), and at yet other times it may be apparent that two systems are conflicting (for example when one is fighting to break an addiction). In fact, some evidence suggests the influence of each system can indeed change over time (Otto et al., 2013a; Lee et al., 2014).

But what drives this change in control? Some work suggests that uncertainty within the model-based and model-free systems may determine that system's influence (Daw et al., 2005; Beierholm et al., 2011; Lee et al., 2014). Such an uncertainty-based arbitration scheme causes decision-making systems that are more confident in their decision to be used, while other systems which are unsure as to the optimal action have less or no control of the agent during that decision. However, it is unclear whether behavioral correlates of deliberation (such as VTE) or procedural learning (such as behavioral stereotypy) correspond to uncertainty within the model-based or model-free systems.

In Daw et al. (2005), the authors use approximate Bayesian versions of the model-based and model-free reinforcement learning algorithms, which are able to express uncertainty in their estimate of the value of taking an action in a given state. This is because the approximate Bayesian versions of the algorithms represent the expected value of each state-action pair as a probability distribution across possible values, instead of by a single scalar estimate, like the non-Bayesian versions of the reinforcement learning algorithms. The uncertainty within a given system is then the variance of that probability distribution of the value of a given state-action pair. If the distribution is wide, then the algorithm is less certain as to the value of the state-action pair, while if the distribution is sharp and the variance is low, the algorithm is highly certain as to the value of taking that action in that state. It is important to note that this form of uncertainty refers to the uncertainty in the estimate of the value of individual state-action pairs, and not to the uncertainty as to which of several competing state-action pairs has the higher value.

We adapted for rats a task which has often been used to dissociate model-based from model-free decision-making in humans. In this chapter, we discuss the task and how we have adapted it for rats, and evaluate rat behavior on our version of the task. We also investigate how choice behavior of rats on our version of the task reflects model-based and model-free influence, and link that behavior to the more traditionally rodent deliberative and procedural behaviors like VTE. We also evaluate what role uncertainty in each of the model-based and model-free algorithms may play in the arbitration between those two decision-making systems.

2.2 The Two-Step Task

To investigate the extent to which rodent behavior can be explained by model-free and model-based influences, how the influence of each algorithm corresponds to habitual and deliberative behavior, and to elucidate how arbitration between these two systems may occur, we adapted for rats the two-step task previously designed for humans which is able to dissociate model-based from model-free decisions (Daw et al., 2011). Rodents are an ideal model species for studying the relationship between information representation in the brain with model-based and model-free algorithms, because large ensembles of single cells can be recorded simultaneously while rodents perform decision-making tasks. Ensemble recordings are an invasive method which result in far more precise measurements of what the brain is doing than, say, magnetic resonance imaging. However, because of its invasiveness, this method is obviously unethical to perform on human subjects, and rodents provide a more cost-effective solution than nonhuman primates.

2.2.1 The original two-step task for humans

The human two-step task (Daw et al. (2011), see Figure 2.1) consists of a sequence of two choices: C1 (choosing between A vs. B) and then C2 (choosing between C vs. D) or C3 (choosing between E vs. F). Choosing option A in C1 usually (but not always!) leads to choice C2, while choosing option B in C1 usually leads to choice C3. Choosing C vs. D (in C2) or E vs. F (in C3) leads to probabilistically-delivered reward, with different probabilities at C, D, E, and F. Another important feature of the two-step task is that the reward probabilities drift slowly over time, so the subject is constantly trying to find the best option and should not simply settle on one option, but can use observations of reward as a signal that the option is a good one to return to (at least for a while).

This human version of the two-step task is able to dissociate model-based from model-free decisions because it creates conditions where the two decision-making algorithms make different choices, mostly on laps following a rare transition (e.g. choosing A at C1 leads to C3, a choice between E and F). This is because the model-based

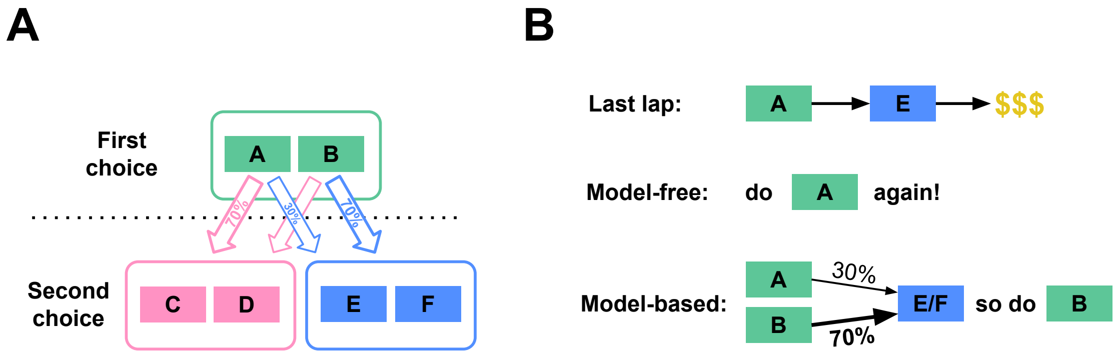

Figure 2.1: The two-step task for humans. (A) State structure of the task. A first choice between two options leads probabilistically to one of two second-stage choices. Each of the four second-stage choices have some cost of reward associated with them, and those costs change over the course of the session. (B) This task dissociates model-based from model-free choices. When an agent receives reward after a rare transition, the model-free system is more likely to repeat the first-stage choice which lead to that reward, while the model-based system is more likely to take the opposite first-stage action on the next lap. Figure from Hasz and Redish (2018).

algorithm stores information about the relation between states (specifically, the state transition probabilities), while the model-free algorithm does not store information about relations between states (and so does not use the transition probabilities for valuation).

To illustrate this difference, suppose a subject chooses A at C1, experiences a rare transition and is presented with C3 (a choice between E and F), chooses E at the second choice, and receives a large reward (Figure 2.1). A model-free agent would be more likely to repeat the choice at C1 (choice A), because model-free learning algorithms reinforce actions which have led to reward in the past, without taking into account relations between states. However, the world model of the model-based algorithm stores relations between states, and so has access to the fact that choosing B at C1 is more likely to lead to the C3 choice, where E can then be chosen. Therefore, the model-based algorithm would be more likely to choose B at C1 in this scenario, while the model-free algorithm would be more likely to choose A. In general on this task, model-based and model-free agents value the two choices at C1 slightly differently.

2.2.2 Our spatial two-step task for rats

Our version of the two-step task for rats was a spatial maze with two sequential left/right choice points (or "stages"), which corresponded to the two choice stages in the human task (Figure 2.2). The second choice (C2/C3) was the same physical location for both the C/D and E/F choices, but an audiovisual cue at the second choice point informed animals whether they were in the C2 or C3 context. Choosing left (A) at the first choice led to C2 80% of the time, and to C3 20% of the time. Like the human task, those probabilities were reversed after choosing right (B) at the first choice point. After choosing left (C or E) or right (D or F) at the second choice point, rats were rewarded with food pellets. While the cost of reward in the human task was the probability of receiving a reward at all, we used delay to food delivery as the cost: high delay to food delivery corresponded to high cost rewards, while low delays corresponded to low cost rewards. Like the human task, these delays varied between C, D, E, and F. The delays were initialized randomly between 1 and 30s, and changed slowly over the course of a session according to a Gaussian random walk with a standard deviation of 1s/lap.

To indicate to the animal which second-stage context they were in, we presented auditory and visual cues after the first choice was made. The auditory cue was a beep pattern unique to each second stage, and the visual cue was white-on-black lines or circles (depending on the second stage) displayed on three monitors around the second choice point. From the pellet dispensers on either side of the maze, there were return hallways to the start of the maze. There was another pellet dispenser at the

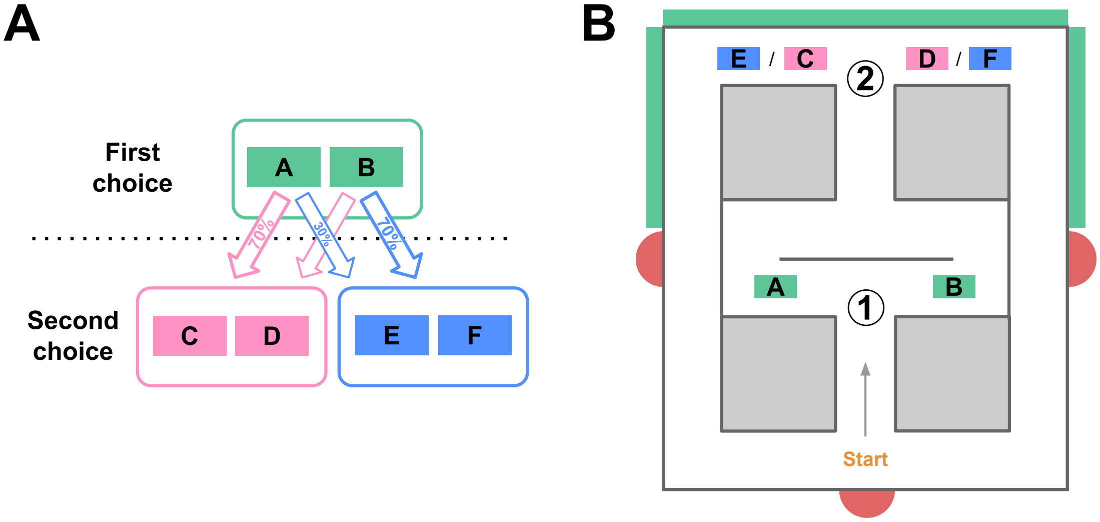

Figure 2.2: The two-step task for rats. (A) State structure of the task is identical to that of the two-step task for humans. (B) The spatial version of the two-step task for rats. An initial Left/Right choice point (labeled "1", corresponding to the first choice in A), leads to a second-stage choice (labeled "2"). Which of the two second stage choices is currently presented is indicated by an audio cue, and by a visual cue on monitors (green boxes on outside of maze). Rats then wait some amount of time before receiving food reward at feeder sites (red semicircles). Figure from Hasz and Redish (2018).

start of the maze, where rats received one pellet per lap. Four one-way servo-actuated doors were used to prevent the rats from moving backwards through the maze: one on either side of the first choice-point, and one just before entry into the reward offer zone. The maze was constructed using LEGO walls and a canvas floor. Rats were allowed to freely run the task for the duration of sessions which lasted 45 min, and earned their food for the day while running the task ($\sim 10 - 15$ g).

Animal behavior on the task was captured with a video camera placed above the maze. Custom Matlab software determined animal position from the video on-line; controlled delays and monitors; controlled pellet dispensers and the one-way doors via an Arduino, and recorded animal trajectory through the maze along with task events. Custom Matlab software was written to track animal head positions from video off-line.

There were three phases of task training, each lasting 8d. For the first, there was no delay to food delivery, no second-stage auditory or visual cues, and one option was blocked at each choice point, leaving only one possible path through the maze. Choices were blocked on sequential days such that all four paths through the maze (LL, LR, RL, RR) were sampled equally. That is, the right side of the first choice point and the right side of the second choice was blocked on the first training day, then on the second training day the right side of the first choice and the left side of the second choice were blocked, and so on. Eight pellets were dispensed at the two feeder sites per reward on the first day of training, and the number of pellets decreased by 1 pellet every two days for the duration of the training phase. A single pellet per lap was dispensed at the rear feeder site.

For the second training phase, there were still no second-stage auditory or visual cues, and one of the first-stage options was blocked, but both second-stage choices were left open. Delay to food was set randomly between 1 and 10s on the first day of second phase training, and the maximum delay increased by 2s/day for the duration of the training phase. The delay values were allowed to change over the course of the session according to the same Gaussian random walk used in the full task (but not allowed to increase above the maximum delay for the day). Four pellets were dispensed at each feeder site for the first four days of this training phase, and three pellets for the last four days.

The third training phase was 8d of the full task, with no choices blocked, a maximum delay of 30s, and two pellets per feeder site.

One drawback to evaluating place and response strategies on traditional rodent tasks, or even identifying VTE at single choice points, is that these behaviors are measured on a per-trial basis, and so it is impossible to determine how the decision-making strategies might evolve over the course of single trials. Therefore, using traditional rodent tasks it is difficult to evaluate whether animals deliberate over single decisions independently, or whether they enter deliberative or habitual modes over the course of an entire trial and make all decisions therein using that policy. A further possibility is that deliberation at the initiation of a trial instigates an epoch of procedural control, which remains for the rest of the trial. Essentially, the question is on tasks where each trial consists of a complex sequence of decisions, whether rats deliberate at each choice, or whether they "plan out" their entire trial from the beginning and follow that plan procedurally. The two-step task provides a method to access this question: by having multiple decisions per trial, we are able to evaluate how rats' decision strategies evolve over the course of single trials.

Furthermore, on traditional rodent tasks, the transition from deliberative to habitual control is usually quantified only as a function of time. For example, by measuring the strength of place/response strategies across trial within a session, or session within a training regimen (Packard and McGaugh, 1996). Assuming automation increases as a function of an animal's experience with that specific action chain, then behavioral stereotypy should increase not only with time, but specifically with the number of actions or choices that the animal has performed. Again the two-step task allows us to evaluate whether this is true without depending solely on time: because the reward values change over time, sometimes rats will experience negligible differences in reward contingencies from lap to lap, in which case they will in theory strengthen the action chain leading to reward, while on other laps the reward value will have changed significantly, and we can measure how the strength of their procedural automation differs in these cases.

But most importantly, the two-step task enables us to measure model-based and

A) Sessions per Rat

Rat	Number of Sessions
1	48
2	50
3	50
4	50
5	53
6	53
7	53
Total	357

B) Laps per Rat

Rat	Number of Laps
1	3313
2	3602
3	4079
4	3610
5	3594
6	3805
7	4478
Total	26481

Table 2.1: The number of sessions and laps run by each rat

model-free influences on rat choice behavior, while simultaneously measuring deliberative and habitual behaviors, and allows for neural activity and representations to be related to model-based and model-free influence.

2.3 Rat Behavior on the Two-Step Task

Rat behavior on the spatial two-step task was collected from seven male Brown Norway rats aged 6-15 months for at least 48 sessions each (357 sessions in total, Table 2.1A). Before behavioral training, rats were handled daily for 7d to accustom them to the experimenter, then acclimated for 7d to eat the food pellets delivered during the task (45-mg sucrose pellets), and finally trained to run through the one-way doors on a separate maze for 7d. Rats were housed on a 12-hr light-dark cycle, and behavioral sessions were run at the same time daily. Rats were food restricted to encourage them to run the task, and maintained weight at >80% of their free-feeding weight. Water was always available in their home cage. All experimental and animal care procedures complied with US National Institutes of Health guidelines for animal care and were approved by the Institutional Animal Care and Use Committee at the University of Minnesota.

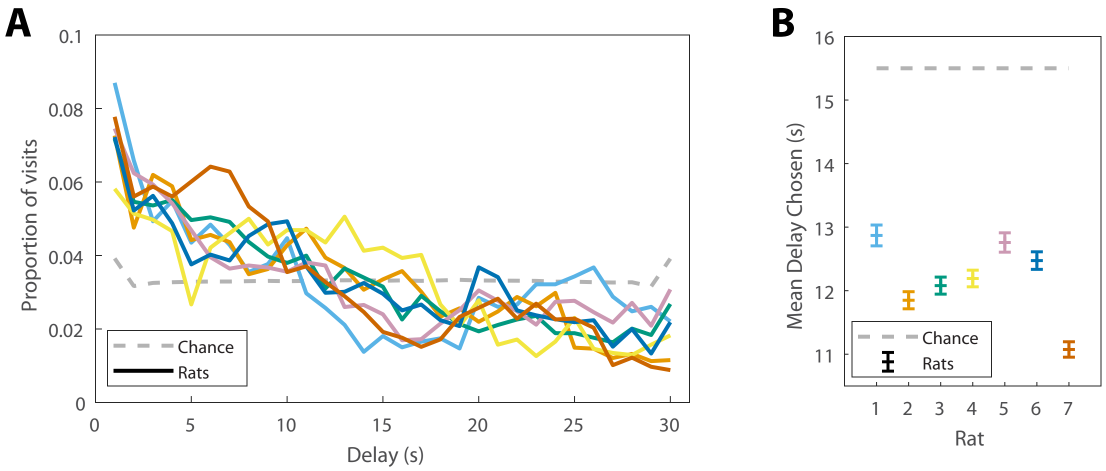

Figure 2.3: Rats displayed a preference for low-delay feeders on the spatial two-step task. (A) The proportion of delays experienced by the rats (colored solid lines, each line is one rat), as compared to the proportions of delays which would be expected by visiting feeders randomly. (B) The mean delay experienced by the rats (+/- SEM) as compared to the mean delay which would be expected by visiting feeders randomly (generated by a model-free simulation run with learning rates at 0). Delays have been aggregated over all sessions from a given rat. Figure from Hasz and Redish (2018).

2.3.1 Rats made choices which led to short delays

Rats ran an average of 74.2±19.6 laps per session on the spatial version of the two-step task (Table 2.1B). Not surprisingly, rats preferred reward offers with a low delay to food delivery (Figure 2.3). We ran simulations of agents which made random choices on the two-step task to determine the delays which would be expected by visiting feeders randomly. That is, at each of the two choice points, the simulated agents had an equal probability of choosing left vs. right. We simulated $10,000$ sessions of this random-choice agent on the two step task, using 74 laps per session (the average length of the rats' sessions).

All rats had a visibly stronger preference for low delays than did the random choice agent simulations (Figure 2.3). Mean delays experienced by the rats were significantly less than the mean delay experienced by the random-choice simulations (two-sided

Wilcoxon signed rank test, $N_{rats} = 7$, $p = 0.0156$, rat delays were 3.31 seconds lower on average than simulation delays). This indicates that rats were able to learn the task, by making decisions which led to lower-delay outcomes.

2.3.2 Rats displayed VTE at choice points

Vicarious trial and error (VTE) is a behavioral correlate of deliberation in rats, characterized by a pause at a choice point, while simultaneously swinging of the head back and forth between potential paths as if deliberating over which path to take (Muenzinger and Gentry, 1931; Tolman, 1939; Redish, 2016). We used LogIdPhi, a measure of pausing and head-turning, to measure VTE (Papale et al., 2012). The LogIdPhi for a given choice point pass corresponds to the angular acceleration of the rat's head, integrated over a pass through the choice point. Therefore, it captures both how long the rat hesitates at the choice point, and how quickly the rat's head is changing direction. When x and y are the position of the rat's head,

$$\text{LogIdPhi} = \log\left(\int_{zone\ entry}^{zone\ exit} \left| \frac{\delta}{\delta t}\ \text{atan2}\left(\frac{\delta y}{\delta t}, \frac{\delta x}{\delta t}\right) \right| \delta t\right) \tag{2.1}$$

On a very small proportion of choice point passes, we were unable to compute VTE due to a momentary lag in the rat position tracking system. At the first choice point, this occurred on 13 laps (0.049% of laps). At the second choice point, this occurred on 10 laps (0.038% of laps). We excluded these laps from our analysis.

We found that on our spatial two-step task, rats displayed varying levels of LogIdPhi at the first choice point (Figure 2.4). There was a clear bimodal distribution of LogIdPhi at the first choice point, where one peak with lesser LogIdPhi values corresponded to laps where VTE did not occcur (Figures 2.4A and 2.4C) and the other peak with greater LogIdPhi values corresponded to laps where VTE occurred (Fig-

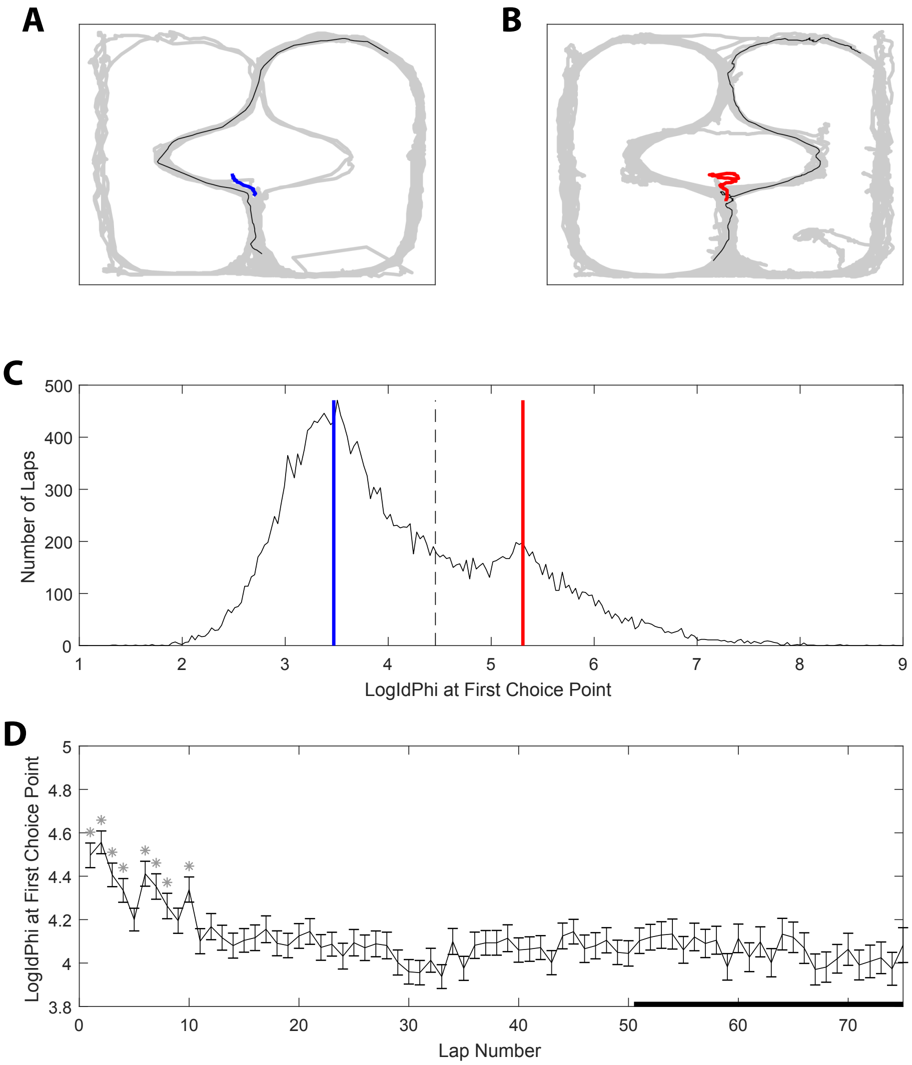

Figure 2.4: Vicarious trial and error (VTE) at the first choice point. (A) An example of a pass through the first choice point without VTE, and (B) an example of VTE at the first choice point. Grey line is rat body position over the whole session, black line is rat body position on example lap, and red or blue lines are rat head position at the first choice point on the example lap. (C) Distribution of LogIdPhi values at the first choice point over all laps, sessions, and rats. Blue line corresponds to LogIdPhi value at the first choice point in the example lap shown in A, and the red line to the example lap shown in B. Dashed line is the VTE/non-VTE threshold (see methods). (D) LogIdPhi over the course of a session. Error bars indicate SEM. Stars indicate laps for which LogIdPhi was significantly greater than that of laps 51 and greater. Data has been aggregated over rats ($N = 357$, the total number of sessions). Error bars show SEM. Figure from Hasz and Redish (2018).

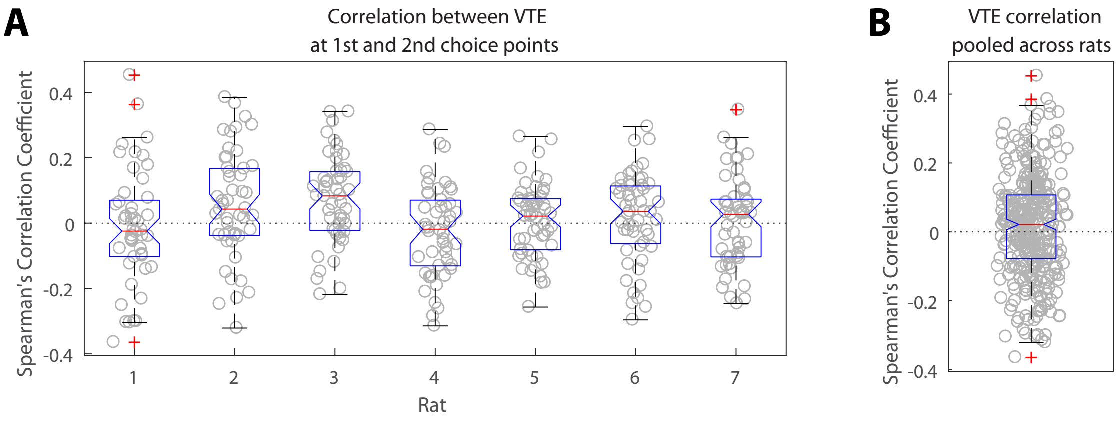

Figure 2.5: Correlation between VTE at the first and second choice points. (A) Correlation coefficients per session for each rat individually. (B) Correlation coefficients per session pooled across rats. Figure from Hasz and Redish (2018).

ures 2.4B and 2.4C). The amount of VTE was greater at the beginning of a session (Figure 2.4D). When comparing each lap to laps > 50, there was significantly more VTE at the first choice point for 8 of the first 10 laps. However, there was not significantly more VTE on laps 10-50 than on laps > 50 (Figure 2.4D, Wilcoxon rank sum test, Bonferroni corrected for multiple comparisons, with pre-correction threshold of $p < 0.05$).

2.3.3 VTE was correlated between within-lap decisions

The two-step task contains two left/right choice points within a single trial, which enabled us to evaluate how deliberative behavior changed over the course of each trial. We found that the amount of VTE at the first and second choice points on a given lap were correlated (Figure 2.5, the median Spearman's correlation coefficient between LogIdPhi at the first and second choice points within a session was greater than 0, two-sided Wilcoxon signed rank test, $N_{sessions} = 357$, $p = 0.0337$, median $\rho = 0.0215$), although this correlation was very slight. Considered individually, 2 individual rats

Rat	Median ρ	p
1	-0.0243	0.406
2	0.0430	0.0267
3	0.0834	0.00109
4	-0.0185	0.178
5	0.0216	0.661
6	0.0359	0.198
7	0.0270	0.982

Table 2.2: Spearman's correlations between VTE at choice point 1 and choice point 2 for each rat. Shown are the median correlation coefficients (over sessions from that rat) and the p-value of a Two-sided Wilcoxon signed rank test.

showed significant positive correlations, while no rats showed significant negative correlations (Figure 2.5A and Table 2.2).

We also fit a mixed model to VTE at the two choice points, to determine if there was a correlation between the amount of VTE at each choice point even while accounting for rat- and session-specific differences in VTE. Specifically, the model tried to predict zLogIdPhi (the z-scored LogIdPhi) at the second choice point from zLogIdPhi at the first choice point on that same lap. The z-scored LogIdPhi was simply z-scored across all rats, laps, and sessions for the first and second choice points independently. These models included subject and session as random effects; that is, they allowed levels of VTE to vary across subjects and sessions, but not in a totally independent way. Our model included a fixed intercept, a fixed effect of transition type on the current lap, a fixed effect of transition type on the previous lap, a per-subject random effect, and a per-session random effect.

$$\text{zLogIdPhi}_{2,i} \sim \mathcal{N}(\beta_0 + \beta_{VTE} \times \text{zLogIdPhi}_{1,i} + R_r + S_s,\ \sigma_e) \tag{2.2}$$

where R and S are the random effects coefficients for rat and session, respectively.

$$\begin{aligned} R &\sim \mathcal{N}(0, \sigma_r) \\ S &\sim \mathcal{N}(0, \sigma_s) \end{aligned} \tag{2.3}$$

where

- $zLogIdPhi_{2,i}$ is the z-scored LogIdPhi value at the 2nd choice point on lap i,
- $zLogIdPhi_{1,i}$ is the z-scored LogIdPhi value at the 1st choice point on lap i,
- β_0 is the fixed intercept of the model (baseline LogIdPhi),
- β_{VTE} is the standardized coefficient (a parameter which captures the relationship between the amount of VTE at the two choice points),
- R_r is rat r's random effect (or adjustment coefficient), which accounts for the possibility that some rats have different baseline levels of LogIdPhi,
- S_S is session s's random effect, which accounts for the possibility that rats have different baseline levels of LogIdPhi on different sessions,
- σ_r and σ_s are the standard deviations of per-rat (R) and per-session (S) random effects, respectively,
- σ_e is the standard deviation of the error, and
- $\mathcal{N}(\mu, \sigma)$ represents a normal distribution centered at μ with standard deviation σ.

Using this mixed model, we found a significant positive correlation between the levels of VTE at the two choice points on single laps (Table 2.3). This suggests that instead of deliberating at each single choice independently, rats may have entered a deliberative mode for entire trials, where then each individual decision within that trial was made using the deliberative system.

Mixed Model of the correlation between VTE at the two choice points

Parameter	2.5%	Estimate	97.5%	t-statistic	DF	p
β	0.0570	0.0685	0.0801	11.7	26457	2.65×10^{-31}
σ_r	0.129	0.225	0.392			
σ_s	0.341	0.369	0.401			
σ_ϵ	0.896	0.904	0.912			

Table 2.3: Mixed Model of the correlation between VTE at the two choice points

2.3.4 Path stereotypy increased over the course of the session

In contrast to vicarious trial and error, path stereotypy is a behavioral correlate of procedural decision-making (Packard and McGaugh, 1996; Jog et al., 1999; Schmitzer-Torbert and Redish, 2002; van der Meer et al., 2012; Smith and Graybiel, 2013; Schmidt et al., 2013). To measure path stereotypy, we used the inverse of the mean distance between the path on a given lap and all other paths during the same session of the same type (LL, LR, RL, or RR), re-sampled in time (Schmitzer-Torbert and Redish, 2002). This resulted in a value which was larger when paths were more stereotyped (similar to the average path), and smaller for irregular paths through the maze. When a lap was the only lap of its type in a session, we could not calculate path stereotypy (with no similar paths for which to compute the mean distance), and so we excluded such laps from our analysis. These laps made up a very small proportion of the total data (0.66%).

The stereotypy of rats' paths also varied on our task (Figure 2.6). Unlike VTE, there was a unimodal distribution of path stereotypy, where some laps were less stereotyped (Figures 2.6A and 2.6C) and other laps were more stereotyped (Figures 2.6B and 2.6C). Also unlike VTE, path stereotypy increased steadily over the course of a session, with 48 of the first 50 laps being significantly less stereotyped than laps greater than 50 (Figure 2.6D, Wilcoxon rank sum test, Bonferroni corrected for multiple comparisons, with pre-correction threshold of $p < 0.05$).

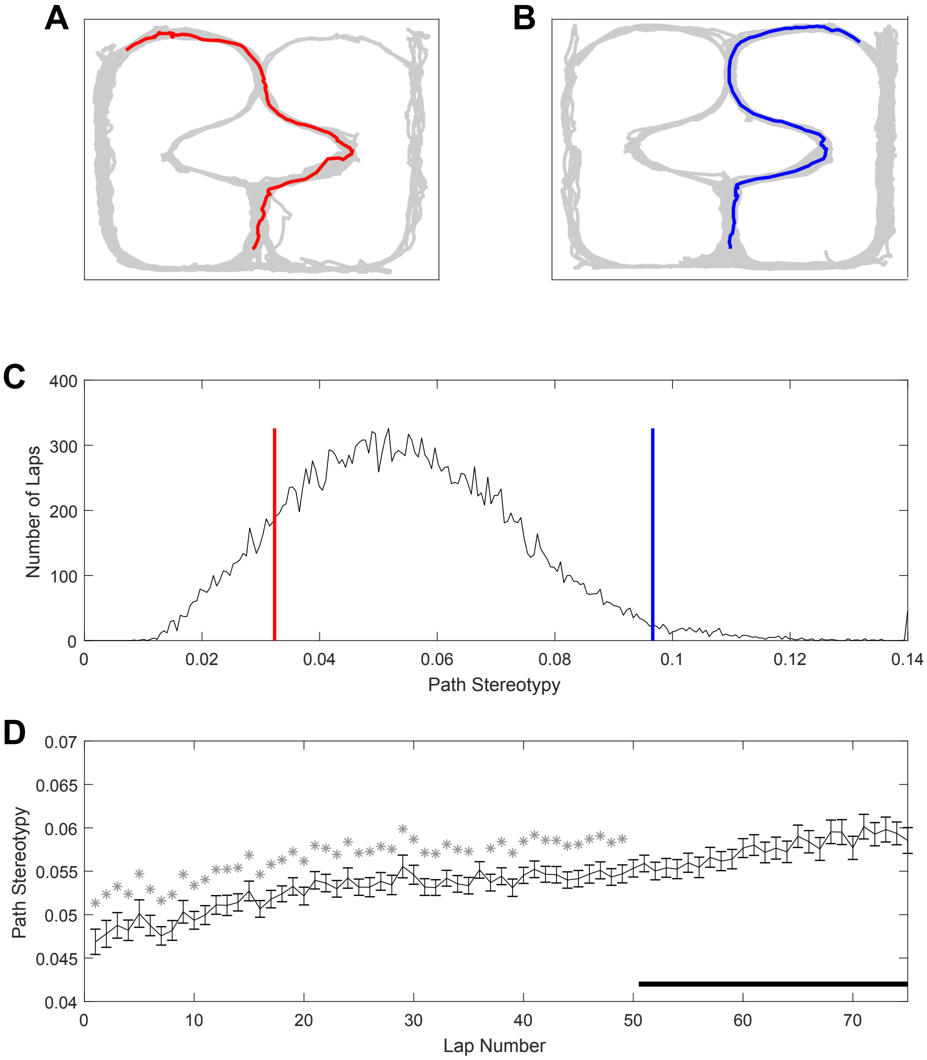

Figure 2.6: Path stereotypy on the spatial two-step task. (A) An irregular, non-stereotyped path, and (B) an example of a highly stereotyped path. The grey line is rat body position over the whole session, and colored lines are the rat body position on the example lap. (C) Distribution of negative log deviation from the average path over all laps, sessions, and rats. Red line corresponds to the log deviation value of the example lap shown in A, blue line to the example lap shown in B. (D) Negative log deviation from the average path over the course of a session. Stars indicate laps for which average path deviation was significantly greater than that of laps 51 and greater. Data has been aggregated over rats ($N = 357$, the total number of sessions). Error bars show SEM. Figure from Hasz and Redish (2018).

2.3.5 VTE and stereotypy were related to choice repeats

Previous rodent research has found that animals transition from displaying deliberative behavior to stereotyped behavior over the course of a session, or with experience on a task. If this shift towards stereotyped behavior is due to procedural learning, then a decrease in deliberative behavior and a corresponding increase in stereotyped behavior should also be apparent as a function of the number of repeated choices an animal makes, and not only as a function of time within the session or training regimen. For the two-step task, we defined a "repeated choice" to be when a rat made the same choice at both the first and second choice points as on the previous lap.

We found that VTE at the first choice point was negatively correlated with the number of repeated choices rats made on the two-step task (Figure 2.7A, E, and H; the per-rat median Spearman's correlation coefficient between LogIdPhi at the first choice point and the number of choice repeats was less than 0, two-sided Wilcoxon signed rank test, $N_{rats} = 7$, $p = 0.0156$, median $\rho = -0.205$). On the other hand, path stereotypy was positively correlated with the number of repeated choices (Figure 2.7D, G, and J; the per-rat median Spearman's correlation coefficient between path stereotypy and the number of choice repeats was greater than 0, two-sided Wilcoxon signed rank test, $N_{rats} = 7$, $p = 0.0156$, median $\rho = 0.274$). We found no significant correlation between VTE at the second choice point and the number of choice repeats (Figure 2.7B, F, and I; the per-rat median Spearman's correlation coefficient between LogIdPhi at the second choice point and the number of choice repeats was not significantly different from 0, two-sided Wilcoxon signed rank test, $N_{rats} = 7$, $p = 0.156$, median $\rho = -0.0730$).

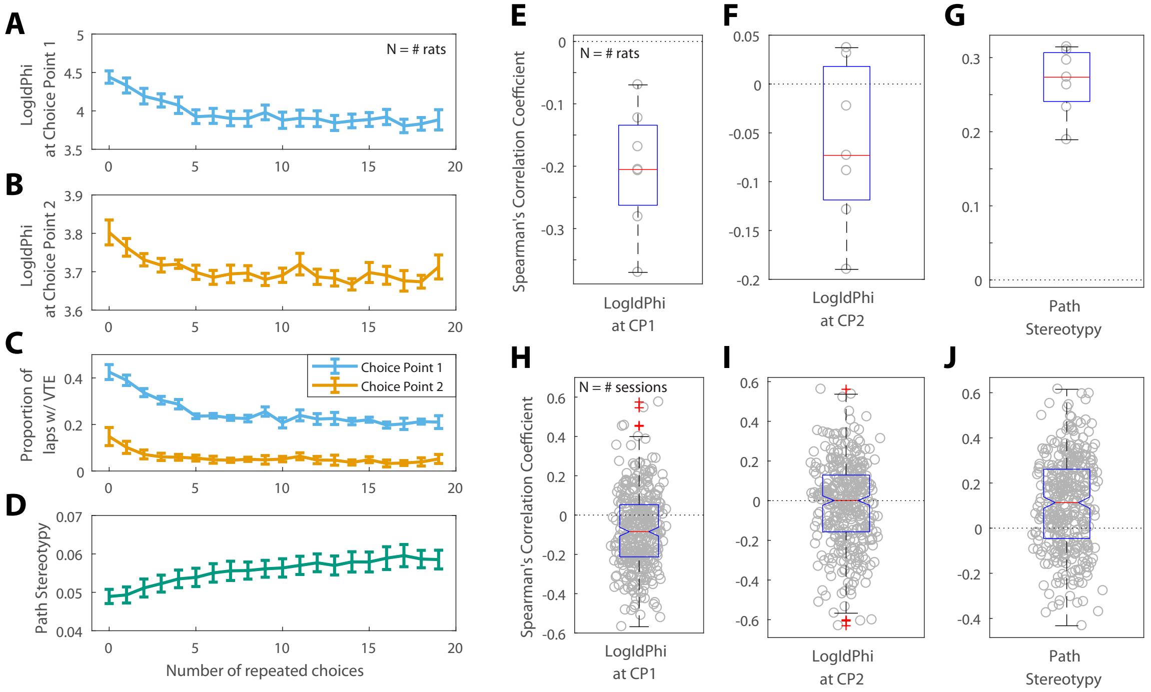

Figure 2.7: VTE and Path Stereotypy as a function of the number of repeated choices. Raw levels of VTE at the first (A) and second (B) choice points, the ratio of laps on which rats showed VTE (C), and path stereotypy (D) as a function of choice repeats. For A-D, error bars show mean +/- SEM with $N = 7$, the number of rats. (E-F) Per-rat correlation coefficients between the number of repeated choices and VTE at the first choice point (E), second choice point (F), and path stereotypy (G). (H-J) Per-session correlation coefficients between the number of repeated choices and VTE at the first choice point (H), second choice point (I), and path stereotypy (J). Figure from Hasz and Redish (2018).

2.3.6 VTE at the second choice was related to transition type

However, the amount of VTE at the second choice point did change depending on whether the transition on that lap was common or rare. We fit linear mixed models for VTE at the first choice point, for VTE at the second choice point, and for path stereotypy, with transition type (common or rare) on the current and previous laps as fixed variables, and rat and session as random variables. These models included subject and session as random effects; that is, they allowed levels of VTE or path stereotypy to vary across subjects and sessions, but not in a totally independent way.

Our model included a fixed intercept, a fixed effect of transition type on the current lap, a fixed effect of transition type on the previous lap, a per-subject random effect, and a per-session random effect.

$$Y_i \sim \mathcal{N}(\beta_0 + Tt_i + T_P t_{i-1} + R_r + S_s,\ \sigma_e) \tag{2.4}$$

where R and S are the random effects coefficients for rat and session, respectively.

$$\begin{aligned} R &\sim \mathcal{N}(0, \sigma_r) \\ S &\sim \mathcal{N}(0, \sigma_s) \end{aligned} \tag{2.5}$$

and

- Y_i is the LogIdPhi value at the first choice point on lap i (or the LogIdPhi value at the second choice point on lap i for the second choice point model, or the path stereotypy value on lap i for the path stereotypy model) ,
- β_0 is the intercept of the model (baseline LogIdPhi or path stereotypy value),
- T is the parameter capturing the fixed effect of rare transitions on the current lap,
- t_i is an indicator variable which is 0 when there was a common transition on lap i, and 1 when there was a rare transition on lap i,
- T_P is the parameter capturing the fixed effect of a rare transition on the previous lap,
- t_{i-1} is an indicator variable which is 0 when there was a common transition on lap $i-1$, and 1 when there was a rare transition on lap $i-1$,

Mixed Model for LogIdPhi at Choice Point 1

Parameter	2.5%	Estimate	97.5%	t-statistic	DF	p
β_0	3.979	4.168	4.357	43.28	26106	$< 10^{-100}$
T	-0.01987	0.009811	0.03949	0.6479	26106	0.517
T_P	-0.004464	0.02525	0.05496	1.666	26106	0.0958
σ_r	0.1424	0.2476	0.4307			
σ_s	0.3734	0.4049	0.4390			
σ_ϵ	0.9622	0.9706	0.9790			

Table 2.4: Mixed model of VTE at the first choice point, with transition type on the current lap and previous lap as fixed effects, and rat and session as random effects. The 2.5% column indicates the lower bound of the 95% confidence interval, and the 97.5% column indicates the upper bound of the 95% confidence interval. DF = degrees of freedom.

- R_r is rat r's random effect (or adjustment coefficient), which accounts for the possibility that some rats have different baseline levels of LogIdPhi or path stereotypy,
- S_S is session s's random effect, which accounts for the possibility that rats have different baseline LogIdPhi or path stereotypy values on different sessions,
- σ_r and σ_s are the standard deviations of per-rat (R) and per-session (S) random effects, respectively,
- σ_e is the standard deviation of the error, and
- $\mathcal{N}(\mu, \sigma)$ represents a normal distribution centered at μ with standard deviation σ.

Laps which were the first in a session were not used in this analysis, as the transition type of the previous (nonexistent) lap was undefined. The degrees of freedom in the mixed model for path stereotypy were different from the degrees of freedom in the mixed models for VTE because on some laps path stereotypy could not be calculated

Mixed Model for LogIdPhi at Choice Point 2

Parameter	2.5%	Estimate	97.5%	t-statistic	DF	p
β_0	3.696	3.726	3.756	243.8	26106	$< 10^{-100}$
T	0.01528	0.02556	0.03584	4.874	26106	1.100×10^{-06}
T_P	-0.009982	0.0003090	0.01060	0.05892	26106	0.9530
σ_r	0.01959	0.03678	0.06906			
σ_s	0.1006	0.1095	0.1191			
σ_ϵ	0.3334	0.3363	0.3392			

Table 2.5: Mixed model of VTE at the second choice by transition type

Mixed Model for Path Stereotypy

Parameter	2.5%	Estimate	97.5%	t-statistic	DF	p
β_0	0.04815	0.05263	0.05712	23.00	25965	$< 10^{-100}$
T	-0.001344	-0.0008540	-0.0003650	-3.420	25965	6.276×10^{-4}
T_P	-0.001006	-0.0005160	-0.00002600	-2.064	25965	0.03900
σ_r	0.0033665	0.0058732	0.010247			
σ_s	0.0095065	0.010264	0.011082			
σ_ϵ	0.015813	0.015951	0.01609			

Table 2.6: Mixed model of path stereotypy by transition type

(when a lap was the only lap of that type in a session). Also the degrees of freedom in the mixed models for VTE are different here than for the mixed model used between VTE at the two choice points, because this model does not include laps which were the first in a session.

There was a significant increase in the amount of VTE at the second choice point following a rare transition (Table 2.5). VTE at the first choice point on the lap following a transition did not significantly differ between common and rare transitions (Table 2.4). Path stereotypy on a given lap, however, was significantly decreased when there was a rare transition either on that lap or on the preceding lap (Table 2.6).

2.3.7 VTE at the first choice was driven by multiple factors

To determine what may have been driving VTE at the first choice point, we fit a mixed model of VTE at the first choice point, with random effects of rat and session, and with fixed effects of the transition on the previous lap, whether the rat repeated its previous choice, and the delay on the previous lap. This model included subject and session as random effects, a fixed intercept, a fixed effect of transition type on the previous lap, a fixed effect of delay experienced on the previous lap, and a fixed effect of choice repetition (whether the previous choice was repeated or not).

$$Y_i \sim \mathcal{N}(\beta_0 + T_P t_{i-1} + D_P d_{i-1} + C c_i + R_r + S_s,\ \sigma_e) \tag{2.6}$$

where R and S are the random effects coefficients for rat and session, respectively.

$$\begin{aligned} R &\sim \mathcal{N}(0, \sigma_r) \\ S &\sim \mathcal{N}(0, \sigma_s) \end{aligned} \tag{2.7}$$

where

- Y_i is the LogIdPhi value at the first choice point on lap i
- β_0 is the intercept of the model (baseline LogIdPhi value),
- T_P is the parameter capturing the fixed effect of a rare transition on the previous lap,
- t_{i-1} is an indicator variable which is 0 when there was a common transition on lap $i-1$, and 1 when there was a rare transition on lap $i-1$,
- D_P is the parameter capturing the fixed effect of the delay on the previous lap,
- d_{i-1} is the delay in seconds on lap $i-1$,

- C is the parameter capturing the fixed effect of choice repetition,
- c_i in an indicator variable which is 0 when the rat did not repeat its choice on lap i, and 1 when it did,
- R_r is rat r's random effect (or adjustment coefficient), which accounts for the possibility that some rats have different baseline levels of LogIdPhi or path stereotypy,
- S_S is session s's random effect, which accounts for the possibility that rats have different baseline LogIdPhi or path stereotypy values on different sessions,
- σ_r and σ_s are the standard deviations of per-rat (R) and per-session (S) random effects, respectively,
- σ_e is the standard deviation of the error, and
- $\mathcal{N}(\mu, \sigma)$ represents a normal distribution centered at μ with standard deviation σ.

We found that VTE at the first choice point was driven by a complex interaction between these three factors (Table 2.7). Confirming our previous results, there was not a detectable main effect of the transition on the previous lap, and there was a significant negative correlation between VTE at the first choice point and repeated choices. There was also a significant positive correlation between delay on the previous lap and VTE at the first choice point. Several of the interaction terms and the three-way interaction were also significant. Taken together, this suggests that VTE at the first choice point reflects a deliberative process, where the interaction between many task variables are being taken into account, instead of simply being driven by a single task variable such as transition.

Mixed Model for LogIdPhi at Choice Point 1

Parameter	2.5%	Estimate	97.5%	t-statistic	DF	p
β_0	4.142	4.328	4.514	45.59	26110	$< 10^{-100}$
T_P	-0.2222	-0.1086	0.005064	-1.873	26110	0.0611
C	-0.5217	-0.4608	-0.3999	-14.83	26110	1.60×10^{-49}
D_P	0.002559	0.00567	0.00878	3.573	26110	0.000354
T_P*C	0.08108	0.2126	0.3441	3.168	26110	0.00153
T_P*D_P	-0.0007716	0.005783	0.01234	1.729	26110	0.0838
C*D_P	0.008183	0.0119	0.01562	6.272	26110	3.62×10^{-10}
T_P*C*D_P	-0.02229	-0.01439	-0.006489	-3.57	26110	0.000358
σ_r	0.1358	0.2353	0.4077			
σ_s	0.3282	0.3566	0.3874			
σ_ϵ	0.9514	0.9597	0.9678			

Table 2.7: Mixed model of VTE at the first choice point. Transition type on the previous lap, delay on the previous lap, and whether the rat repeated its choice or not are fixed effects, and rat and session are random effects. A*B indicates an interaction term between A and B. The 2.5% column indicates the lower bound of the 95% confidence interval, and the 97.5% column indicates the upper bound of the 95% confidence interval. DF = degrees of freedom.

These results indicate that VTE at the first and second choice points may have been partially driven by different factors. VTE at the first choice point occurred more often when rats had just switched to a new choice pattern and interactions between task variables, but was not detectably affected by the transition on the previous lap alone. On the other hand, VTE at the second choice point occurred more often after an unexpected transition, but was not detectably affected by choice repetitions. We hypothesize that VTE at the first choice point arises more as a result of some deliberative process, which in theory also decreases with the number of repeated choices. Conversely, we hypothesize that VTE at the second choice point, when not being driven by a deliberative mode, arises more as a result of the interruption of a procedural process, which may lead to deliberation, because it is influenced more strongly by unexpected transitions in the middle of a lap than by a change in choice patterns.

The correlation between VTE at the two choice points may seem inconsistent with our interpretation that VTE at the second choice point is driven by an interruption of a procedural process. However, we do not believe that VTE at the second choice point is being driven entirely by such interruptions. Rather, we would hypothesize that VTE at the second choice point likely co-occurs with VTE at the first choice point when rats are in a deliberative mode, and that VTE at the second choice point is only primarily driven by rare transitions when rats are in a procedural mode and the unexpected transition interrupts their stereotyped behavior.

2.4 Rats Display a Mix of Model-based and Model-free Decision-Making

Do rat choices on the two-step task reflect influences of model-based and model-free decision making algorithms? First, we'll explain in detail how these algorithms actually work. Then, we'll compare the behavior of simulations of model-based and model-free agents on the two-step task to choice patterns of the rats. Finally, we'll fit these reinforcement learning algorithms to rat behavior in order to determine what kinds of models best explain rat behavior on the two step task.

Each algorithm computed the expected value (or Q-value) of taking an action a, in any given state, s. Our model of the two-step task included only two possible actions in any state ("go left" or "go right"), and only three states: the first choice point (C1, a choice between A and B), and the two possible second choice points (C2, a choice between C and D; and C3, a choice between E and F, see diagram in Figure 2.2).

The next 3 subsections explain how each algorithm computes the expected value (or Q-value) of taking an action a, in any given state, s. The section after that describes how the likelihood is computed for each algorithm from that algorithm's Q-values. This "likelihood" is the probability that the algorithm, with a given set of values for its parameters, would make the same choices we observed the rats make on the two-step task. Then, we compare simulations of these agents to rat behavior on the two-step task, and use Bayesian inference and model comparison to determine which model is most likely to explain rat behavior.

2.4.1 The model-free algorithm

For the model-free algorithm, we used the SARSA(λ) temporal difference learning algorithm (Rummery and Niranjan, 1994), as was used in Daw et al. (2011). This algorithm learns the expected value (Q_{MF}) of taking a given action a, in any given state s, by updating the Q values according to the delta rule:

$$Q_{MF}(s_{i,t}, a_{i,t}) = Q_{MF}(s_{i,t}, a_{i,t}) + \alpha_i \delta_{i,t} \tag{2.8}$$

where $s_{i,t}$ is the state on trial t at stage i, and $a_{i,t}$ is the action taken in that state on that trial. α_i is the learning rate for stage i. There were only two stages on the two-step task: decisions at the first stage (C1) used α_1, and decisions at the second stage (C2 or C3, see Figure 2.2) used α_2. The reward prediction error, $\delta_{i,t}$, was the difference between expected and experienced reward on trial t at stage i:

$$\delta_{i,t} = r_{i,t} + Q_{MF}(s_{i+1,t}, a_{i+1,t}) - Q_{MF}(s_{i,t}, a_{i,t}) \tag{2.9}$$

where $r_{i,t}$ is the reward experienced at stage i of trial t. For the first stage reward, we defined $r_{1,t} = 0$, because rats did not receive reward between the first and second choice points. For the second stage rewards, we defined the reward as the opposite of the cost:

$$r_{2,t} = d_{max} - d_{2,t} \tag{2.10}$$

where d_{max} is the maximum possible delay to food on our task (30 seconds), and $d_{2,t}$ is the delay experienced on trial t (and explicit delays only occurred after a choice at stage 2). This assumes that rats are aware of the maximum delay, which we believe is a valid assumption, because rats were trained extensively on the task before the experiment began. We also defined a third "virtual" state, where $Q_{MF}(s_{3,t}, a_{3,t}) = 0$, because there is no further reward in a trial following food delivery. The algorithm updates the first-stage state-action value based on the eligibility trace parameter and second-stage reward prediction error at the end of each trial:

$$Q_{MF}(s_{1,t}, a_{1,t}) = Q_{MF}(s_{1,t}, a_{1,t}) + \alpha_1 \lambda \delta_{2,t} \tag{2.11}$$

Note that with the SARSA algorithm the update for $Q_{MF}(s_{1,t}, a_{1,t})$ occurs twice per trial: once after the first-stage choice (where the α_1 learning rate is used), and again after the end of the trial according to the eligibility trace parameter, λ (where a learning rate of $(\alpha_1 \lambda)$ is used, as in equation 2.11.

2.4.2 The model-based algorithm

The model-based algorithm updates the state-action values of the second-stage states ($Q(a_{2,t}, s_{2,t})$) in exactly the same way as the model-free system. However for the first-stage state action values, instead of updating them according to the delta rule, the model-based algorithm takes into account the transition probabilities and the best option at either second stage, and computes the first-stage action values at decision time by:

$$\begin{aligned} Q_{MB}(s_A, a_t) = \quad & p(s_B|s_A, a_t)\ \max\nolimits_{a' \in \{a_A, a_B\}} Q_{MF}(s_B, a') \\ & + p(s_C|s_A, a_t)\ \max\nolimits_{a' \in \{a_A, a_B\}} Q_{MF}(s_C, a') \end{aligned} \tag{2.12}$$

where s_A is the first-stage state, s_B is one of the two second-stage states, s_C is the other second-stage state, and a_t is an action taken at the first stage of trial t. $p(s_X|s_Y, a_t)$ is the transition probability from state s_Y to s_X after taking action a_t at s_Y. Because the rats were trained on the two-step task for over three weeks before we started collecting the data to which these models were fit, we assumed the rats had learned the transition probabilities by the end of training, and so our model did not include the learning of the transition probabilities. Therefore $p(s_X|s_Y, a_t)$ was set to either 0.8 for a common transition or 0.2 for a rare transition.

2.4.3 The constant-weight hybrid algorithm

This algorithm values actions according to some constant weight between the model-based and model-free algorithm values. Essentially, the constant-weight hybrid algorithm "runs" both the model-free and model-based algorithms simultaneously, and then computes the value (Q_{CW}) of taking some action a in some state s as the weighted average between the state-action values of the model-free and model-based systems:

$$Q_{CW}(s, a) = wQ_{MB}(s, a) + (1 - w)Q_{MF}(s, a) \tag{2.13}$$

where w is a free parameter which determines the weighting between the model-based and model-free systems. If $w = 1$ then the algorithm is purely model-based, and if $w = 0$ then the algorithm is purely model-free. The model-based and model-free algorithms within the constant-weight hybrid algorithm are assumed to share parameters, as in Daw et al. (2011).

However, note that this assumption may not actually be true: for example, the procedural system is thought to have a far slower learning rate than the deliberative system. It would be interesting for further work to examine if more complex models which allow the two systems to have independent parameters better explain rat or human behavior. Here, however, we stick to the parameter-sharing version of the constant weight model, in order to most closely match the models used in Daw et al. (2011).

2.4.4 Computing the likelihood of each algorithm

To transform each algorithm's valuations of different state-action pairs (each algorithm's Q-values) into probabilities that the algorithm would make the same choice as the rats did at stage i of trial t (we denote this probability by $p(a_{i,t} = a|s_{i,t})$), we used a softmax for each algorithm, in the same way as in Daw et al. (2011):

$$p(a_{i,t} = a|s_{i,t}) = \frac{\exp(\beta_i[Q(s_{i,t}, a) + p \times rep(a)])}{\sum_{a'} \exp(\beta_i[Q(s_{i,t}, a') + p \times rep(a')])} \tag{2.14}$$

where β_i is an inverse temperature parameter that controls how stochastic the models' choices are at each choice point, and the sum in the denominator sums over all available actions, a'. As $\beta_i \to 0$, the choices become purely random, and as $\beta_i \to \infty$, the probability of choosing the action with the largest Q value approaches 1. We used independent β_i parameters for each stage of the task, and the i index of β_i corresponds to the stage. There were only two stages on the two-step task. Decisions at the first stage (C1) used β_1, and decisions at the second stage (C2 or C3, see Figure 2.2) used β_2.

The p parameter accounts for an inclination to repeat the same action taken on the last lap ($p > 0$), or to switch to the opposite action ($p < 0$), regardless of expected action values. $rep(a)$ was a function which evaluated to 1 if the rat repeated its action, that is, performed action a at that stage on the previous lap (stage i, trial $t - 1$), and 0 if it chose a different action. Therefore if the p parameter was positive, the algorithm was more likely to repeat the previous choice, and if it was negative, the algorithm was more likely to switch (choose the opposite choice from the previous trial). The purpose of this p parameter was to capture perseveration behavior.

We initialized all Q values to the mean reward value at the beginning of each session. The log likelihood of observing rat choices across all N_s sessions given an algorithm was then computed by summing the log likelihood of each choice for each stage, lap, session, and rat:

$$\log(p(\text{data}|\theta)) = \sum_{d=1}^{N_d} \sum_{t=1}^{N_t} \sum_{i=1}^{N_i} \log\left(p(a_{i,t} = a|s_{i,t})\right) \tag{2.15}$$

where θ is the set of all parameters for a given algorithm, N_i is the number of choice stages in each trial t (for our task this is always 2: the first choice point, C1, and the second choice point, C2 or C3, see Figure 2.2), N_t is the number of trials in a given session (or "day") d, and N_d is the total number of sessions across all rats.

2.4.5 Rat behavior compared to algorithm simulations

We ran simulations of model-free and model-based agents on the two step task, and compared the choice patterns of the simulated agents to those of the rats. The model-free and model-based simulations were generated by 10,000 simulated sessions of model-free or model-based agents with 74 trials per session (the average number of trials per session run by the rats). Parameters used for the simulations were

$\alpha_1, \alpha_2 = 0.5,\ \beta_1, \beta_2 = 3,\ p = 0.3,\ \lambda = 0$ for both the model-free and model-based agents.

On the two-step task, our simulated model-free agents were more likely to repeat first-stage choices which led to low-delay (low-cost) rewards than those which led to high-delay (high-cost) rewards, even if this reward occurred after a rare transition (Figure 2.8A). However, model-based agents were more likely to show the opposite pattern after rare transitions – that is, they are less likely to repeat first-stage choices which led to low-cost rewards than those which led to high-cost rewards after rare transitions (Figure 2.8B). The choice patterns of rats on the two-step task appeared neither purely model-based nor purely model-free, suggesting a mix of model-based and model-free behavior (Figure 2.8C), consistent with behavior seen in human subjects (Gläscher et al., 2010; Daw et al., 2011).

2.4.6 Bayesian reinforcement learning model fits

To more rigorously evaluate model-based or model-free influences on rat choices, we fit model-based and model-free algorithms to rat choices on the two-step task. We also considered the constant-weight hybrid algorithm where choices were made according to some fixed weight between model-based and model-free influence. Specifically, we performed Bayesian inference with these models using Markov chain Monte Carlo (MCMC) in Stan (Carpenter et al., 2017), and the Python programming language interface to Stan, PyStan (Stan Development Team, 2017), to generate model parameter posterior distributions so that we could perform model comparison and inference of the parameter values (Kruschke, 2014). Stan is a platform for Bayesian statistical modeling (`http://mc-stan.org`), in which models can be written using a simple modeling language, and Stan performs MCMC sampling resulting in model and pa-

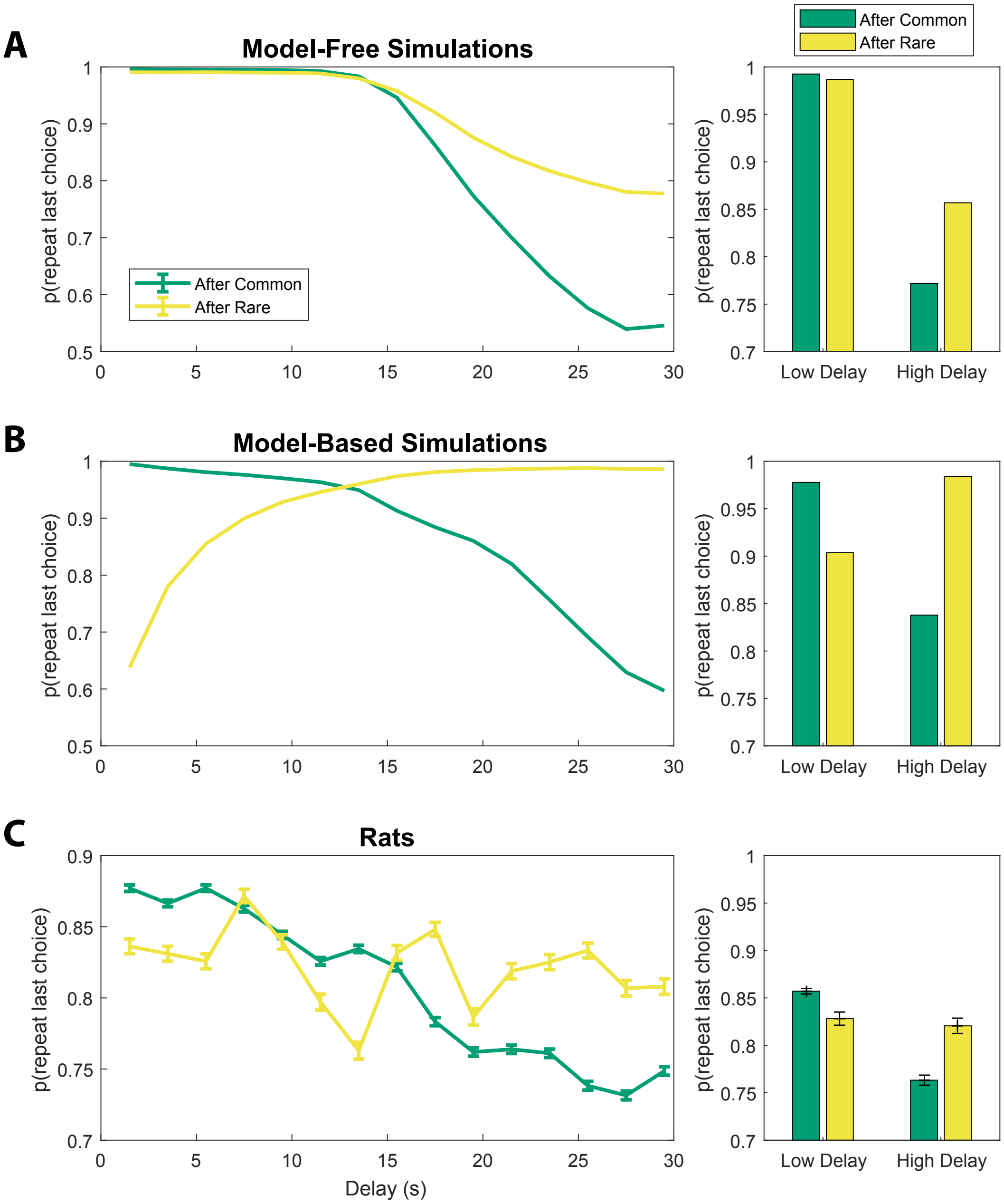

Figure 2.8: First-stage choice repetition by delay for (A) model-free and (B) model-based reinforcement learning simulations. Data has been aggregated over simulated sessions. Error bars were omitted from A and B because SEM of the simulations was negligible. (C) Rats show features of both model-based and model-free behavior. Data has been aggregated over rats and sessions. Error bars show SEM with $N =$ the total number of laps with a given delay. Figure from Hasz and Redish (2018).

rameter posterior probabilities. This allowed us to perform Bayesian inference as to the values of model parameters, and model comparison using DIC scores.

We used vaguely informative priors for the Bayesian fits in Stan. Across all models, the priors used were:

Parameter	**Prior**
α_1	Beta distribution with $\alpha = 1.2, \beta = 1.2$
α_2	Beta distribution with $\alpha = 1.2, \beta = 1.2$
λ	Beta distribution with $\alpha = 1.2, \beta = 1.2$
β_1	Exponential distribution with $\lambda = 0.5$
β_2	Exponential distribution with $\lambda = 0.5$
p	Normal distribution with $\mu = 0, \sigma = 10$
w	Beta distribution with $\alpha = 1.2, \beta = 1.2$

Table 2.8: Priors used for reinforcement learning models

Each algorithm was fit in PyStan with 5 chains per algorithm, and 10,000 iterations per chain (5000 warm-up and 5000 sampling). Chains which took longer than 96 hours to run were aborted and re-started. We used pooled (non-hierarchical) models, such that the same parameter was used for each rat.

Note that each of the three models had a unique number of parameters, with the constant-weight algorithm having the most:

Algorithm	**Number of Parameters**	**List of parameters**
Model-free	6	$\alpha_1, \alpha_2, \lambda, \beta_1, \beta_2$, and p
Model-based	4	$\alpha_2, \beta_1, \beta_2$, and p
Constant-weight	7	$\alpha_1, \alpha_2, \lambda, \beta_1, \beta_2, p$, and w

Table 2.9: Number of parameters per reinforcement learning model

Using naive model comparison methods, like comparing model likelihoods, could cause models with more parameters to be deemed more likely due to overfitting. Therefore, we used Deviance Information Criterion (DIC) scores to select the most likely of these three algorithms (Spiegelhalter et al., 2002). DIC allows a more fair comparison of models with different numbers of parameters by penalizing models which have a higher effective number of parameters. It is also well-suited for use with models whose posterior distributions have been computed via MCMC, which is the method we used. Given MCMC samples of parameter values θ (a vector of parameter values), we compute the DIC score by:

$$DIC = D(\bar{\theta}) + 2p_D \tag{2.16}$$

where the effective number of parameters (p_D) is computed by:

$$p_D = \bar{D} - D(\bar{\theta}) \tag{2.17}$$

$\bar{D}$ is the average of the deviance, $D(\theta)$, over all the MCMC samples of θ:

$$\bar{D} = \frac{1}{N_{samples}} \sum_{i=1}^{N_{samples}} D(\theta_i) \tag{2.18}$$

$D(\bar{\theta})$ is the deviance evaluated at the average of the MCMC samples of θ:

$$D(\bar{\theta}) = D\left(\frac{1}{N_{samples}} \sum_{i=1}^{N_{samples}} \theta_i\right) \tag{2.19}$$

and the deviance is computed by:

$$D(\theta) = -2\log(p(\text{data}|\theta)) \tag{2.20}$$

where $\log(p(\text{data}|\theta_i))$ is the algorithm likelihood, as computed above (in section 2.4.4), given parameters θ for a MCMC sample. The deviance is technically $D(\theta) = -2\log(p(\text{data}|\theta)) + C$, but C is a constant which cancels out when comparing different models. Algorithms are compared based on their DIC scores, where models with lower DIC scores are more likely to explain the data. Differences in DIC scores greater than 7 suggest the algorithm with the higher DIC score has "considerably less support" (Spiegelhalter et al., 2002) than the algorithm with the lower DIC score.

The purely model-based algorithm was more likely than the purely model-free algorithm to explain rat choices on the two-step task (DIC score difference of 94, Tables 2.10, 2.11, and 2.13). In tables 2.10-2.13, MAP: maximum a posteriori parameter estimate; Mean: mean of the MCMC samples for that parameter; Std: standard deviation of the MCMC samples for that parameter; DIC score: deviance information criterion for that model; Log Post.: mean log posterior probability. However, the constant-weight hybrid algorithm was more likely than the purely model-based algorithm to explain rat choices on the two-step task (DIC score difference of 69, Tables 2.11, 2.12, and 2.13). The fact that the constant-weight hybrid algorithm had a far lower DIC score suggests that rat choices on the two-step task were driven by some combination of model-based and model-free decision making, and were not driven by either the model-based or model-free system alone. This is consistent with many human studies which find that human choices on the two-step task display both model-based and model-free influences (Gläscher et al., 2010; Daw et al., 2011; Wunderlich et al., 2012; Otto et al., 2013b,a; Doll et al., 2016).

2.4.7 Discussion

Our findings are consistent with previous work in humans which finds that hybrid algorithms are more likely to explain behavior than model-based algorithms alone, and that the weights in these hybrid algorithms favor model-free decision-making (Daw et al., 2011; Voon et al., 2015), though see Simon and Daw (2011) and Gillan

Model-free

Parameter	MAP	Mean	Std
α_1	0.0710	0.0739	0.0120
α_2	0.00165	0.00170	0.000551
β_1	3.44	3.73	1.20
β_2	3.64	3.93	1.28
p	0.380	0.387	0.120
λ	0.00200	0.00171	0.00140
DIC score:	51515	Log Post.:	-25832

Table 2.10: Model-free algorithm fit to rat behavior

Model-based

Parameter	MAP	Mean	Std
α_2	0.000933	0.000920	0.000240
β_1	7.29	7.87	1.98
β_2	6.39	6.90	1.74
p	0.177	0.174	0.0451
DIC score:	51421	Log Post.:	-25741

Table 2.11: Model-free algorithm fit to rat behavior

et al. (2015). However, some work in rodents on the two-step task finds that rodent choices are primarily, but not necessarily exclusively, model-based or "planning-driven" (Miller et al., 2013; Akam et al., 2013; Miller et al., 2014, 2017). This discrepancy could have been caused by any of several factors, but we suspect differences in how we implemented the two-step task for rodents was the main contributor.

There were some specific differences between our version of the two-step task and that used by others. Unlike the human version of the two-step task (Daw et al., 2011) and other rodent adaptations (Miller et al., 2017), we used delay to reward delivery as the cost, instead of the probability of reward delivery. We also implemented the full version of the two-step task, with costs which changed according to a random walk, and no second stage choice cue. The more simplified version used in rodents by Miller et al. (2017) had costs which switched between blocks of trials but stayed

Constant Weight

Parameter	MAP	Mean	Std
α_1	0.0371	0.0360	0.0196
α_2	0.00121	0.00129	0.000360
β_1	6.16	6.55	1.84
β_2	4.96	5.01	1.38
p	0.207	0.211	0.0593
λ	0.00144	0.00190	0.00207
w	0.675	0.647	0.0795
DIC score:	51352	Log Post.:	-25735

Table 2.12: Constant-weight algorithm fit to rat behavior

Relative DIC scores

Model	Constant Weight	$<$	Model Based	$<$	Model Free
DIC difference	(most likely)	69		94	(least likely)

Table 2.13: DIC scores between reinforcement learning algorithms

constant throughout a block, and had a cued second stage choice.

We found that reinforcement learning models were difficult to fit to rat choices on our task. The number of MCMC iterations required to obtain fits whose chains converged was extremely high ($\sim 10,000$), and attempting to fit multilevel models (models with rat as a mixed effect) only aggravated this problem. Furthermore, the fit learning rates of our reinforcement learning models were suspiciously low (see Tables 2.10 and 2.11). We suspect that the complexity of our version of the two-step task for rodents, along with the use of delay to reward delivery as the cost, prevented the rats from learning the task well enough to employ solely the model-based system, and so relied also on the model-free system in order to make choices on the task. This may explain why we found that a mix of model-based and model-free strategies best explained rat choices on our task.

We also noticed that some rats preferred certain feeders over multiple days, regardless of delay (data not shown here). It could be that Pavlovian decision-making or place preferences also played a role in some rats' choices. This might explain in part the relatively low values of the fit second-stage learning rates (see Tables 2.10 and 2.11). In the current analysis, we chose not to model side biases in order to limit our models to the simplest set of model features which were able to capture model-based vs. model-free choices. However, it would be informative in future work to investigate and model the influences of other decision-making systems in addition to only the model-based and model-free systems.

Hierarchical learning, or "chunking" of action sequences, is thought to occur when multiple actions are chained together and are able to be released as a single action. While action chains are usually thought to be driven by a model-free system, some work suggests that model-based systems are capable of initiating action chains which may appear driven by procedural learning (Dezfouli and Balleine, 2012, 2013; Dezfouli et al., 2014). In future work, it would be interesting to investigate if and how the effects of hierarchical learning on the two-step task affect (or are affected by) arbitration between systems.

Our task used the same two physical locations for the four second-stage end states. Although the task included auditory and visual cues, some rats may have confused the two second-stage end states which shared the same location (for example they may have confused E and C, or D and F, see Figure 2.2B). This may have caused some "bleeding" between the expected values of state-action pairs which led to those states. Any confusion of states in this way would have been an error in situation recognition, and would not necessarily have been occurring in the model-based or model-free systems themselves. Situation recognition is thought to be carried out by a separate system, one not intrinsic to the model-based or model-free systems themselves (Redish et al., 2007; Fuhs and Touretzky, 2007; Gershman et al., 2015). Therefore, any confusion between states would presumably affect both the model-based and model-free systems equally. For this reason we decided not to model any bleeding of state-action values because we were interested only in differences between the model-based and model-free systems.

We adapted the two-step decision task from Daw et al. (2011) for rats in order to study behavioral correlates of model-free and model-based decision-making, but another main advantage of a spatial version of the task is that it can also be used to study neural correlates of model-free and model-based decision-making using electrophysiological techniques in the rodent brain. Representation of state-action pairs and "task-bracketing" in dorsolateral striatum have been hypothesized to initiate action sequences which have been learned procedurally (Jog et al., 1999; Frank, 2011; Regier et al., 2015b). On the other hand, model-based neural activity has been observed in a variety of brain areas including hippocampus, ventral striatum, orbitofrontal cortex, prefrontal cortex, and dorsomedial striatum (Johnson and Redish, 2007; van der Meer et al., 2012; Daw and Dayan, 2014; Wikenheiser and Redish, 2015; Brown et al., 2016), and inactivating the dorsal hippocampus in rats impairs model-based decisions (Miller et al., 2017). The current behavioral analysis assumes that either the model-based or model-free system is used to make a decision, but it would be informative to record from the neural structures implicated in procedural learning and those involved in deliberation in rats as they run the two-step task to determine if and how the two systems operate concurrently. That said, this spatial version of the task was difficult for rats to learn, and further work is required to create a spatial version of the task for rodents which enables both the collection of a large number of trials per session, and allows animals to better learn the task.

Also, Akam et al. (2015) suggest that certain model-free strategies can appear to generate model-based choices on the two-step task. Therefore, if these systems may not be able to be conclusively dissociated based purely on choice patterns, it will be important for further work to investigate neural activity in brain areas thought to drive model-based or model-free decision making in order to truly disentangle the contribution of each system.

By adapting for rats a decision task which is made up of multi-choice trials, we were able to investigate how rats used model-free and model-based choice strategies on the task, along with how the transition from deliberation to procedural automation occurs over the course of single trials, and over the course of sequences of repeated choices. We found that a mixture of model-based and model-free choice strategies was more likely to explain rats' choices on this task than either strategy alone. Furthermore, we found that vicarious trial and error at the two choices within a trial were correlated, which suggests that rats entered deliberative or procedural modes for whole laps. Also, vicarious trial and error at the first choice in a trial corresponded to a complex interaction between task variables and the number of repeated choices, suggesting a deliberative process. Conversely, we found that vicarious trial and error at the second choice in a trial corresponded to unexpected transitions, suggesting it was driven by interruptions in a procedural process which triggered deliberation.

2.5 Uncertainty-based Arbitration between Decision Making Systems

In the previous section we identified that a mixture of model-based and model-free influences appear to drive rat decisions on the two-step task. However, presumably this weighting is not constant. If the hypothesis is correct that deliberation is driven by a model-based mechanism, and procedural behavior by a model-free mechanism, then we would expect that sometimes the model-free system is primarily in control, while at other times the model-based system is primarily in control. This is because rats often display more deliberative behavior early in training (Figure 2.4D) or before making many repeated identical choices (Figure 2.7A), while they display more procedural behavior with extensive experience on a task (Figure 2.6D) or after making many repeated identical choices (Figure 2.7D).

But what drives this change in control? How are multiple decision-making systems within the brain arbitrated between? The animal is only a single agent which obviously is only able to make one single coherent action, so how does the brain decide which of the decision making systems to use, or if each come to a decision independently, how does the brain combine their decisions into a single action plan?

Daw et al. (2005) hypothesize that uncertainty in each system is what decides which system is used. That is, they propose that the system which is more confident in its decision has more control over the animal's or agent's action. In that work, the authors use approximate Bayesian versions of the model-based and model-free reinforcement learning algorithms discussed in section 2.4. These algorithms capture in their estimate of the value of taking a given action in a given state (the Q-values) by representing the Q values as probability distributions, instead of point values as in the previously discussed versions of the algorithms. The uncertainty of a given system at any moment in time is the variance of the distribution representing the expected reward associated with the state-action pair being experienced.

However, this form of uncertainty may not be the only type of uncertainty that is relevant for a decision-making system. The flavor of uncertainty captured by the models used by Daw et al. (2005) express only the uncertainty as to the amount of reward expected from the action which was actually taken by the agent. However, another type of uncertainty would capture the difference in the mean expected rewards obtained from competing actions. A third type of uncertainty would capture both the difference between the mean expected rewards and the variance associated with those estimates. We designed versions of the Bayesian reinforcement learning algorithms which use each of these three types of uncertainty to arbitrate between the model-based and model-free decision making systems.

To elucidate the extent to which model-free and model-based uncertainty predicts which system is used to make a decision, and which type of uncertainty (if any) is most important for arbitration, we fit uncertainty-dependent versions of the reinforcement learning algorithms which used different forms of uncertainty to weight the contributions of the model-based and model-free systems on a decision-by-decision basis.

2.5.1 Bayesian reinforcement learning algorithms

We simulated the approximate Bayesian versions of model-based and model-free reinforcement learning algorithms from Daw et al. (2005), given the same experiences as the rats, in order to compute the uncertainty within each algorithm at each of the rats' decisions. Importantly, the models used to estimate uncertainty – the approximate Bayesian models from Daw et al. (2005) – were separate from the models which were being arbitrated between (the non-Bayesian model-based and model-free reinforcement learning algorithms, discussed in section 2.4). That is, the "uncertainty-dependent algorithm" used the uncertainty of the approximate Bayesian models to determine which of the non-Bayesian algorithms to use to make a choice. We did this so that we could compute uncertainty in as similar a way as possible to the method used in Daw et al. (2005).

We converted the delay to a "reward" value between 0 and 1 in order to match the range of reward values in Daw et al. (2005). We assumed that by the time the experiment began (after $>$ 8 days of training), the rats had learned the maximum (d_{max}) and minimum (0) reward delays, and therefore felt it was valid to convert the delay to a value between 0 and 1. For the approximate Bayesian versions of the reinforcement learning models, we calculated reward, R, such that a reward of 1 corresponded to the lowest possible delay and a reward of 0 corresponded to the highest possible delay:

$$R = \frac{d_{max} - delay}{d_{max}} \tag{2.21}$$

The value of each state-action pair was modeled by a beta distribution, which represents the probability that the reward of a state-action pair takes the value R,

$$R \sim \text{Beta}(\alpha, \beta) \tag{2.22}$$

where α and β are the two shape parameters of the beta distribution. Note that the α and β here refer to the two shape parameters of a beta distribution – *not* to the reinforcement learning rate parameters (α_1 and α_2) or the inverse temperature parameters (β_1 and β_2) as in other sections.

Importantly, we use the quantification of uncertainty from Daw et al. (2005), which uses a beta distribution to model the underlying probability of binary outcomes. The outcomes in our task are not binary, but continuous (delay in seconds). In order to stay as close to the quantification of uncertainty used in Daw et al. (2005), we normalized the continuous-valued delays between 0 and 1 (see above), such that we could use the same quantification of uncertainty as used in Daw et al. (2005).

2.5.2 Bayesian model-free algorithm

For each state-action pair's beta distribution, we used a prior of $\alpha, \beta = 1$. That is, at the beginning of each session, we initialized $\alpha, \beta = 1$ for each state-action pair. This resulted in a uniform distribution between 0 and 1. This differed from Daw et al. (2005), who used a prior of $\alpha, \beta = 0.1$. The prior of 0.1 is used in Daw et al. (2005) because it yields a beta distribution with highest density around 0 and 1, and the authors argue that agents probably initially assume that the result of an action is that there either is reward or there isn't. This makes sense for their task because they use probabilistic rewards: for their subjects, there always is a reward or there isn't. However our task is different in two ways: first, we use non-binary rewards (delay), and second, rats have been trained on our task for 8 days before beginning the experiment, instead of experiencing an experimental session only once, as is often done with human subjects. Delays are initialized randomly at the beginning of each session, so after training the rats should be at least somewhat aware that there is a uniform probability of delay at the start of the task. That is, no one delay is more likely than any other. Therefore, we initialize $\alpha, \beta = 1$ because this results in a beta distribution which is uniform between 0 and 1.

Upon reward delivery, the parameters (α and β) of the beta distribution for the experienced second-stage state-action pair were updated with:

$$\alpha^{MF}_{s',a'} = \alpha^{MF}_{s',a'} + R \tag{2.23}$$

$$\beta^{MF}_{s',a'} = \beta^{MF}_{s',a'} + (1 - R) \tag{2.24}$$

where (s', a') is the state-action pair that was experienced at the second stage (so s'

is either C2 or C3), and R is the amount of reward experienced after taking action a' in state s' (a value between between 0 and 1, see definition above).

The shape parameters of the beta distribution for the experienced first-stage state-action pair were then updated using the mean of the distribution for the experienced second-stage state-action pair,

$$\alpha^{MF}_{C1,a} = \alpha^{MF}_{C1,a} + \mu^{MF}_{s',a'} \tag{2.25}$$

$$\beta^{MF}_{C1,a} = \beta^{MF}_{C1,a} + (1 - \mu^{MF}_{s',a'}) \tag{2.26}$$

where (s', a') is the state-action pair that was experienced at the second stage, $(C1, a)$ is the state-action pair that was experienced at the first stage (at choice C1), and

$$\mu^{MF}_{s',a'} = \frac{\alpha^{MF}_{s',a'}}{\alpha^{MF}_{s',a'} + \beta^{MF}_{s',a'}} \tag{2.27}$$

The mean and variance of a model-free 1st-stage distribution was then

$$\mu^{MF}_{C1,a} = \frac{\alpha^{MF}_{C1,a}}{\alpha^{MF}_{C1,a} + \beta^{MF}_{C1,a}} \tag{2.28}$$

$$(\sigma^2)^{MF}_{C1,a} = \frac{\alpha^{MF}_{C1,a}\beta^{MF}_{C1,a}}{(\alpha^{MF}_{C1,a} + \beta^{MF}_{C1,a})^2(\alpha^{MF}_{C1,a} + \beta^{MF}_{C1,a} + 1)} \tag{2.29}$$

As in Daw et al. (2005), we use a decay factor (γ) which causes the state-action beta distributions to decay toward their priors each timestep. At the end of each trial, we decay each state-action distribution shape parameters by

$$\alpha^{MF} = \alpha^{MF} - \gamma(\alpha^{MF} - (\alpha^{MF})_0) \tag{2.30}$$

and

$$\beta^{MF} = \beta^{MF} - \gamma(\beta^{MF} - (\beta^{MF})_0) \tag{2.31}$$

where $(\alpha^{MF})_0$ and $(\beta^{MF})_0$ are the priors on the α^{MF} and β^{MF} parameters, respectively (1 for both, for all state-action pairs). We use a decay factor of $\gamma = 0.02$, as was used by Daw et al. (2005). This decay approximates a learning rate, in that information learned further in the past is weighted less than information learned more recently.

2.5.3 Bayesian model-based algorithm

The model-based Bayesian reinforcement learning algorithm is similar to the model-free Bayesian reinforcement learning algorithm except it takes transition probabilities into account, in order to compute online the probability of reward for first-stage state-action pairs. As with the model-free Bayesian reinforcement learning algorithm, we used a prior of $\alpha, \beta = 1$ for each state-action pair's beta distribution. For second-stage state-action distributions, the model-based state-action distributions were modeled in the same way as in the model-free algorithm. That is, upon reward delivery, the distribution for the experienced second-stage state-action pair was updated with:

$$\alpha^{MB}_{s',a'} = \alpha^{MB}_{s',a'} + R \tag{2.32}$$

and

$$\beta^{MB}_{s',a'} = \beta^{MB}_{s',a'} + (1 - R) \tag{2.33}$$

In Daw et al. (2005), a Dirichlet distribution was used to model state transition probabilities, but we made the simplifying assumption that the rats had learned the transition probabilities during training phase 3. So, we modeled the first-stage model-based state-action beta distributions by

$$\alpha^{MB}_{C1,a} = \sum_{i \in \{C2,C3\}} p(C1 \rightarrow i|a)\ \alpha^{MB}_{i,a_{max}} \tag{2.34}$$

$$\beta^{MB}_{C1,a} = \sum_{i \in \{C2,C3\}} p(C1 \rightarrow i|a)\ \beta^{MB}_{i,a_{max}} \tag{2.35}$$

where state $C1$ is the first-stage state, and $C2$ and $C3$ are the two second-stage states, and $p(C1 \rightarrow i|a)$ is the probability that performing action a at the first-stage state leads to state i (where i is either $C2$ or $C3$). As with the non-Bayesian version of the model-based algorithm, we assumed the rats had learned the transition probabilities by the end of training, and so we set $p(C1 \rightarrow i|a)$ to either 0.8 for common transitions or 0.2 for rare transitions. a_{max} denotes the apparently best action in the given second-stage state (the action with the highest mean expected reward),

$$a_{max} = \text{argmax}_{x \in \{L,R\}} \mu^{MB}_{i,x} \tag{2.36}$$

The mean and variance of a model-based first-stage distribution was then

$$\mu^{MB}_{C1,a} = \frac{\alpha^{MB}_{C1,a}}{\alpha^{MB}_{C1,a} + \beta^{MB}_{C1,a}} \tag{2.37}$$

$$(\sigma^2)^{MB}_{C1,a} = \frac{\alpha^{MB}_{C1,a}\beta^{MB}_{C1,a}}{(\alpha^{MB}_{C1,a} + \beta^{MB}_{C1,a})^2(\alpha^{MB}_{C1,a} + \beta^{MB}_{C1,a} + 1)} \tag{2.38}$$

We decayed the model-based Bayesian reinforcement learning algorithm's state-action distributions toward their priors in exactly the same way as in the model-free

Bayesian reinforcement learning algorithm (Eqs. 2.30 and 2.31), again using $\gamma = 0.02$, as was used by Daw et al. (2005).

We did not use a step penalty parameter, although it was used in Daw et al. (2005), because our task had only two stages, and so a state was never more than one action removed from a terminal state. This parameter was used in Daw et al. (2005) to penalize the variance of state-action pair beta distributions which had non-terminal successor states (those which led to states which were not the end of a trial).

2.5.4 Value Uncertainty

To quantify uncertainty, Daw et al. (2005) used the variance of the beta distribution representing first-stage state-action pair reward values (Figure 2.9A). We refer to this type of uncertainty as "value uncertainty," because it refers to uncertainty as to the value of a specific state-action pair. So, the value uncertainty of the model-free system on lap i (before making the first-stage decision on that lap) was

$$u^{MF}_{value}(i) = (\sigma^2)^{MF}_{C1,a}(i) \tag{2.39}$$

and the value uncertainty of the model-based system on lap i was

$$u^{MB}_{value}(i) = (\sigma^2)^{MB}_{C1,a}(i) \tag{2.40}$$

where the action a is the action the rat took at the 1st-stage choice on lap i.

2.5.5 Action Uncertainty

However, another conceivable way of formulating uncertainty would be to use the uncertainty as to *which* action to take. That is, uncertainty as to what action in

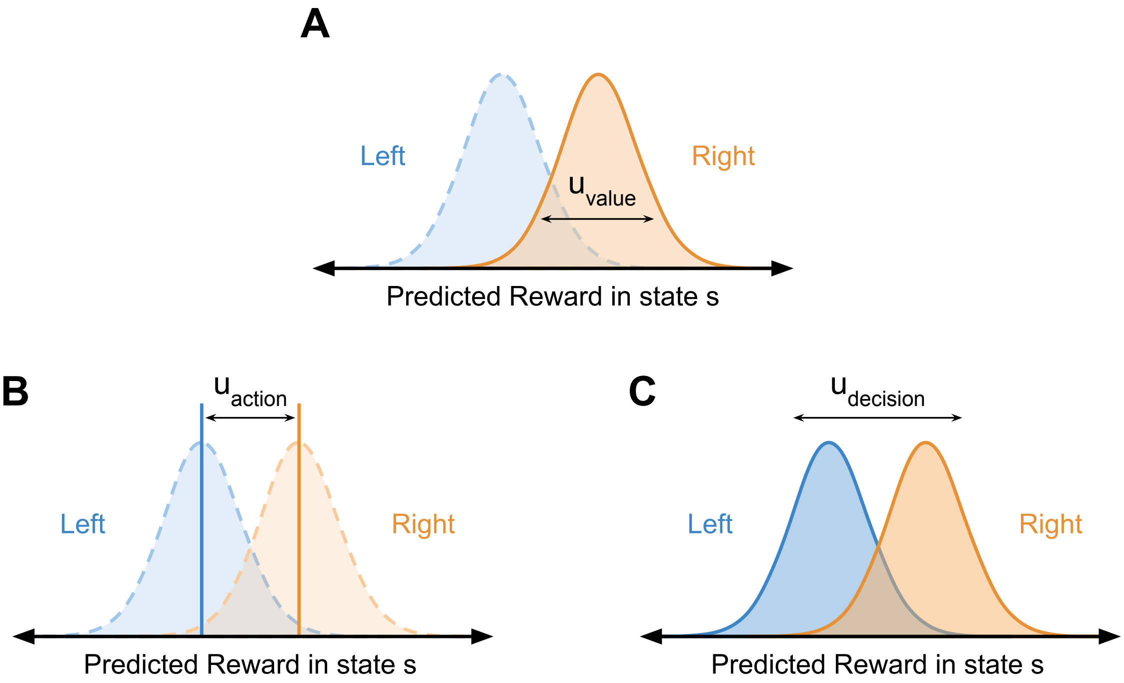
A
Left
Right
u_{value}
Predicted Reward in state s
B
u_{action}
Left
Right
Predicted Reward in state s
C
$u_{decision}$
Left
Right
Predicted Reward in state s

Figure 2.9: Three different types of uncertainty. (A) Value uncertainty, which includes only the uncertainty as to the value estimate of the chosen option. (B) Action uncertainty, which captures only the uncertainty as to which choice has the highest expected value. (C) Decision uncertainty, which captures both mean and variance differences in reward between potential options.

a given state has the highest expected reward (Figure 2.9B). We refer to this type of uncertainty as "action uncertainty," because it refers to uncertainty as to which action has the highest expected reward, instead of to the uncertainty as to the value of a specific state-action pair. To quantify action uncertainty, we used the entropy between the means of the beta distributions representing the expected reward of available actions in a given state. In the two-step task, there were only two actions ever available in any state (left or right). So, the action uncertainty of the model-free system on lap i was

$$u^{MF}_{action}(i) = \mathrm{H}([\mu^{MF}_{C1,L}, \mu^{MF}_{C1,R}]) = -\mu^{MF}_{C1,L} \log_2 \mu^{MF}_{C1,L} - \mu^{MF}_{C1,R} \log_2 \mu^{MF}_{C1,R} \tag{2.41}$$

and the action uncertainty of the model-based system on lap i was

$$u^{MB}_{action}(i) = \mathrm{H}([\mu^{MB}_{C1,L}, \mu^{MB}_{C1,R}]) = -\mu^{MB}_{C1,L} \log_2 \mu^{MB}_{C1,L} - \mu^{MB}_{C1,R} \log_2 \mu^{MB}_{C1,R} \tag{2.42}$$

where $C1$ is the first-stage state, L is the action corresponding to choosing left, and R is the action for choosing right.

2.5.6 Decision Uncertainty

Yet a third way of formulating uncertainty would be to use not just the means or the variances, but to use the entire distribution to compute uncertainty as to what decision to make. Specifically, when the divergence between the reward beta distributions for two available actions is low, uncertainty is high, and vice-versa (Figure 2.9C). We refer to this type of uncertainty as "decision uncertainty," because it refers to uncertainty as to the entire decision when taking into consideration the full expected reward distributions. We quantified decision uncertainty by taking the natural exponential function of the negative symmetrised Kullback-Leibler divergence between the two beta distributions representing the expected reward value of available actions in the 1st-stage state.

So, with

$$P^{MF} = \mathrm{Beta}(\alpha^{MF}_{C1,L}, \beta^{MF}_{C1,L}) \qquad \text{and} \qquad Q^{MF} = \mathrm{Beta}(\alpha^{MF}_{C1,R}, \beta^{MF}_{C1,R}) \tag{2.43}$$

the decision uncertainty of the model-free system on lap i was

$$u^{MF}_{decision}(i) = \exp\left(-D_{KL}(P^{MF}||Q^{MF}) - D_{KL}(Q^{MF}||P^{MF})\right) \tag{2.44}$$

and with

$$P^{MB} = \mathrm{Beta}(\alpha^{MB}_{C1,L}, \beta^{MB}_{C1,L}) \qquad \text{and} \qquad Q^{MB} = \mathrm{Beta}(\alpha^{MB}_{C1,R}, \beta^{MB}_{C1,R}) \tag{2.45}$$

the decision uncertainty of the model-based system on lap i was

$$u^{MB}_{decision}(i) = \exp\left(-D_{KL}(P^{MB}||Q^{MB}) - D_{KL}(Q^{MB}||P^{MB})\right) \tag{2.46}$$

where the Kullback-Leibler divergence (D_{KL}) between two beta distributions was computed with

$$\begin{aligned} D_{KL}(\mathrm{Beta}(\alpha,\beta)||\mathrm{Beta}(\alpha',\beta')) \;\; &= \;\; \ln\left(\tfrac{\mathrm{B}(\alpha',\beta')}{\mathrm{B}(\alpha,\beta)}\right) + (\alpha-\alpha')\psi(\alpha) + (\beta-\beta')\psi(\beta) \\ &\qquad + (\alpha'-\alpha+\beta'-\beta)\psi(\alpha+\beta) \end{aligned} \tag{2.47}$$

where $\mathrm{B}(x)$ is the beta function and $\psi(x)$ is the digamma function.

2.5.7 Uncertainty-based Arbitration

We fit to rat behavior three different uncertainty-based algorithms, each of which used one of the aforementioned three types of uncertainty to arbitrate between the model-based and model-free systems for decision-making. Like the constant-weight algorithm, the uncertainty-based algorithms ran both the model-based and model-free algorithms simultaneously. However, instead of the final state-action values being some constant weighted average between the model-free and model-based state-action values, the uncertainty-based algorithm used the model-based state-action values if the uncertainty of the model-free Bayesian reinforcement learning algorithm was greater than that of the Bayesian model-based reinforcement learning algorithm on a given lap:

$$Q_{UB}(i) = \begin{cases} Q_{MB}(i), & \text{if } u^{MF}(i) > u^{MB}(i) \\ Q_{MF}(i), & \text{otherwise} \end{cases} \tag{2.48}$$

2.5.8 Uncertainty models were difficult to fit to rats' choices

Unfortunately, the fits of the uncertainty-based reinforcement learning algorithms were extremely hard to fit to the rats' choices on the two-step task. The MCMC chains did not converge for any of the three models. Even taking only chains with the best seemingly convergent log likelihoods (chains which seemed to have converged on a single best log likelihood posterior density), the log likelihood was worse for the uncertainty-based models than for even the model-free algorithm, which as seen in section 2.4.6 was otherwise the worst-fitting algorithm. This suggests that while our version of the two-step task was sufficient for determining the contribution model-based and model-free influence overall or on average (section 2.4), it was insufficient for revealing the contributions of model-based and model-free influences on a trial-by-trial basis, which would required for fitting these uncertainty-based models reliably.

As discussed in section 2.4.7, it seemed to be difficult for the rats to learn this version of the two-step task. This could be due to any number of factors, but the most likely culprits seems likely to be the low trial count per session, in combination with the slow speed of the changing delays. A paucity of situations where the delay values were suddenly different from what the rats were expecting (due to the overly slow delay changes) would obstruct our ability to see a difference between the two reinforcement learning algorithms. Those sudden unexpected changes in delay or reward values are the situations where the predictions of the two systems differ, and therefore the only times when our model would be able to parse out the influence of uncertainty on the arbitration between the two models' influences. So, it seems likely that these problems prevented us from accurately capturing trial-by-trial differences in the influence of different decision-making systems, and were therefore unable to asses the influence of uncertainty on the balance between the two decision-making systems.

However, both theoretical work (Daw et al., 2005) and experimental evidence (Beierholm et al., 2011; Lee et al., 2014) suggest that uncertainty within the model-based and model-free systems may indeed determine that system's influence. For future work using this task in rodents, we would suggest using a simplified version of the task (Miller et al., 2013, 2014, 2017), or ensuring the random walk of reward values are fast enough to allow algorithm fits to discern the differences between model-based and model-free influences on behavior.

In order to better study the differences between habitual and deliberative behaviors, as well as the representations of more abstract task features usually associated only with the model-based system, we next developed a different task. The goals of this task were for the task structure to be easier for rats to learn, for the rats to be able to run far more laps within a single session, but for the task to still present rats with a decision-making challenge that would engage both the habitual and deliberative decision-making systems at different times, allowing us to study the differences and dynamics between habitual and deliberative behaviors and neural activity. Therefore, we designed a simpler contingency-switching task, which will be the focus of the next chapter.

Chapter 3

Contingency-Aware Behavior on a Contingency Switching Task

3.1 The Contingency-Switching Task

The two main drawbacks of the two-step task were that rats were unable to run enough laps for us to reliably fit models which captured variables changing on a trial-by-trial basis (like uncertainty), and that the reward values changed too slowly to create drastic differences between habitual and deliberative systems. To address both these problems, we designed a variant of the multiple-T Left/Right/Alternate (MT-LRA) task. This task variant allowed us to study the neural correlates of both the deliberative and habitual systems, but in a way which would be easier for rats to learn, and which had sudden, drastic changes in reward contingencies (unlike the two-step task, which had slowly drifting changes in those contingencies).

The Multiple-T Left/Right/Alternate (MT-LRA) contingency-switching task was a spatial reversal task where rats were required to adjust their behavioral strategies after uncued rule changes. The maze consisted of several low-cost choice points fol-

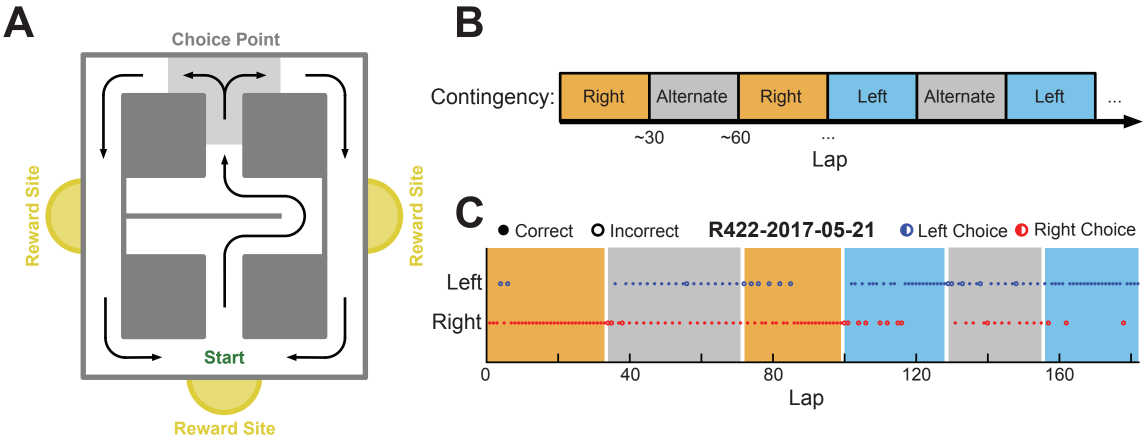

Figure 3.1: The MT-LRA contingency-switching task. (A) The MT-LRA task is a spatial maze with a choice point where rats receive rewards dependent on making choices consistent with the current contingency. (B) Contingencies are presented in blocks of laps lasting 30 ± 5 trials. (C) Example behavioral data from a single session.

lowed by a high-cost choice point between two actions: left or right (Figure 3.1A). The maze was constructed using LEGO blocks on a white canvas. The configuration of the low-cost choice points at the center of the maze was determined by a single wall in the middle of the maze, which switched back and forth from the left to right side randomly each day.

Each lap, if rats chose the action at the high-cost choice point which was consistent with the current contingency, they received one unflavored 45mg food pellet at one of two reward sites on the side of the maze, and an additional food pellet at the rear of the maze. If their choice was inconsistent with the current contingency, no reward was delivered and rats had to circle around to the start of the maze to initiate a new lap. Two different auditory cues also signalled to the rats whether their decision was correct or incorrect: a swept-frequency sinewave "chirp" from 1kHz to 3kHz for correct, and two shorter 1kHz square wave tones for incorrect. The contingency on any given lap was either Left (only left choices at the choice point lead to reward), Right (only right choices), or Alternate (the opposite choice from the previous lap was required for reward). Rats were allowed to run laps freely on the task for one hour each day, and their daily food allowance came only from performing the task. However, rats were fed extra food after running the task if their weight dropped below 80% of their free-feeding weight. This post-feeding occurred after 0 out of 85 experimental sessions, and after 7 out of 212 training sessions (3%).

Rat behavior and neural activity has been studied on previous versions of the MT-LRA task (Gupta et al., 2010; Blumenthal et al., 2011; Steiner and Redish, 2012; Gupta et al., 2012; Powell and Redish, 2014; Regier et al., 2015a; Powell and Redish, 2016). However, these earlier versions of the task included only a single contingency switch halfway through the task session, or no mid-session switch at all (where contingency differences were only between sessions). The main difference between our version of the task and previous iterations is that we modified the task to include multiple uncued contingency changes per session (once per about 30 trials). That is, the contingencies were presented in blocks: every 30 ± 5 laps, the contingency changed randomly to one of the other two contingencies (Figure 3.1B). This allowed us to investigate the reliability of the contingency representations over time, and separate the contributions of any unrelated slow representational changes over time (which could be erroneously construed as contingency representation) from explicit representations of the contingency identity.

How could slow representational changes be misconstrued as the encoding of task contingencies? On the contingency-switching task – and in fact most tasks where there are latent contingencies – those contingencies are presented in blocks of trials. If the contingencies were cued, then the experiment would not so much be studying the ability of animals to use working memory and the deliberative system to perform the task, but simply stimulus-response behaviors. To really access how internal representations of the world (in the form of working memory) are used to make decisions by the deliberative system, we need a task where the contingencies are latent and animals must figure them out for themselves and use their memory of the contingencies to make decisions.

Unfortunately, this presents a problem when we wish to determine if the contingencies are being represented by the brain: if the contingencies are always presented in blocks of trials, then these contingency blocks are synonymous with blocks of time. How then can we determine if the brain is representing the contingency as an abstract rule, or if the brain is simply representing blocks of time – or perhaps other information which is changing over time – in a way perhaps unrelated to contingency?

One way to disentangle the effects of time and contingency representation in the brain is to use a task which has multiple, separated presentations of the same contingency type, and then analyze the reliability of contingency representations across time. This is why we altered the contingency switching task to use more than two contingency blocks per session, to allow us to determine whether contingency representations are stable across multiple presentations of that contingency, or whether the apparent contingency representation is due simply to unrelated change in encoding over time.

Because of the contingency definitions, switches between all contingency types were not identical: switches from L or R blocks to any other type resulted in a 0% reward probability, while switches from A to either L or R resulted in a 50% reward probability (in the case where the new contingency was consistent with the opposite of the choice the rat made on the previous lap, see Table 3.1).

Rat positions on the maze were tracked using a video camera placed above the maze. Custom Matlab software determined animal position from the video, and controlled the state of the task (the current contingency, food pellet release, the presentation of audio cues, etc). The Matlab software also interfaced with an Arduino

Contingency switch type	Reward rate under perseveration
L → R	0%
L → A	0%
R → L	0%
R → A	0%
A → L	50%
A → R	50%

Table 3.1: The reward probability under perseveration (taking actions consistent with the old contingency type) for different contingency switch types.

Uno Rev3 which ran custom software and triggered the release of food pellets from food pellet dispensers.

Rats were trained over the course of four weeks. Starting on the first week, rats were deprived of the freely available food in their home cages, but continued to have free access to water. Rats were handled and offered up to 15g of food pellets each day for half an hour, to train them to eat the food pellets which would be available while performing the LRA task. For the second week, rats performed a simplified version of the task where the contingency was either Left or Right, and the contingency stayed constant throughout each session but changed randomly from session to session. Rats were rewarded with 2 food pellets per feeder at all feeder sites for the second week. For the third week, again there were no within-session contingency switches, but all three contingencies were possible (including Alternate), and only 1 food pellet was delivered at the rear food delivery site. For the final week of training, only 1 food pellet was delivered at all feeder sites, but the task was otherwise the same as during week 3.

After training, rats were given free access to food for at least 3 days, and then surgerized. After 3 days of post-surgery recovery, rats were again food deprived and re-trained for 1-2 weeks on the final training phase of the task (all 3 contingencies possible, but no within-session contingency switches, and 1 pellet per feeder). Finally, rats performed the full version of the task including within-session contingency switches and neural recordings for 2-3 weeks.

3.2 Rat Behavior on the Contingency Switching Task

We ran eight FBNF-1 rats aged 8-14 months at the beginning of behavior on the contingency-switching task (4 male, 4 female), bred from Fischer and Brown Norway rats. Only six of these had usable neural recordings (4 male, 2 female), so in this section we report only behavioral data from those six rats which were used for the neural analyses as well. Rats were housed on a 12 h light-dark cycle, and experimental sessions were conducted at the same time each day during the light phase. All experimental and animal care procedures complied with US National Institutes of Health guidelines for animal care and were approved by the Institutional Animal Care and Use Committee and the University of Minnesota.

Rats ran 137.7 ± 31.7 laps per session (mean ± standard deviation), and encountered 4.1 ± 1.2 contingency switches per session. Rats made correct choices (rewarded choices consistent with the current contingency) on 78 ± 3 percent of laps across all three contingency types, which was significantly more often than chance (4347 correct laps out of 5508, two sided binomial test vs 50%, $p < 10^{-100}$). Rats performed less well on laps during the Alternate contingency (Figure 3.2C), where they made correct choices on only 62.9 ± 9.9 percent of laps, but their performance on laps during the Alternate contingency was still significantly better than chance (1231 correct laps out of 1874, two sided binomial test vs 50%, $p = 1.4 \times 10^{-42}$). Rats did not show

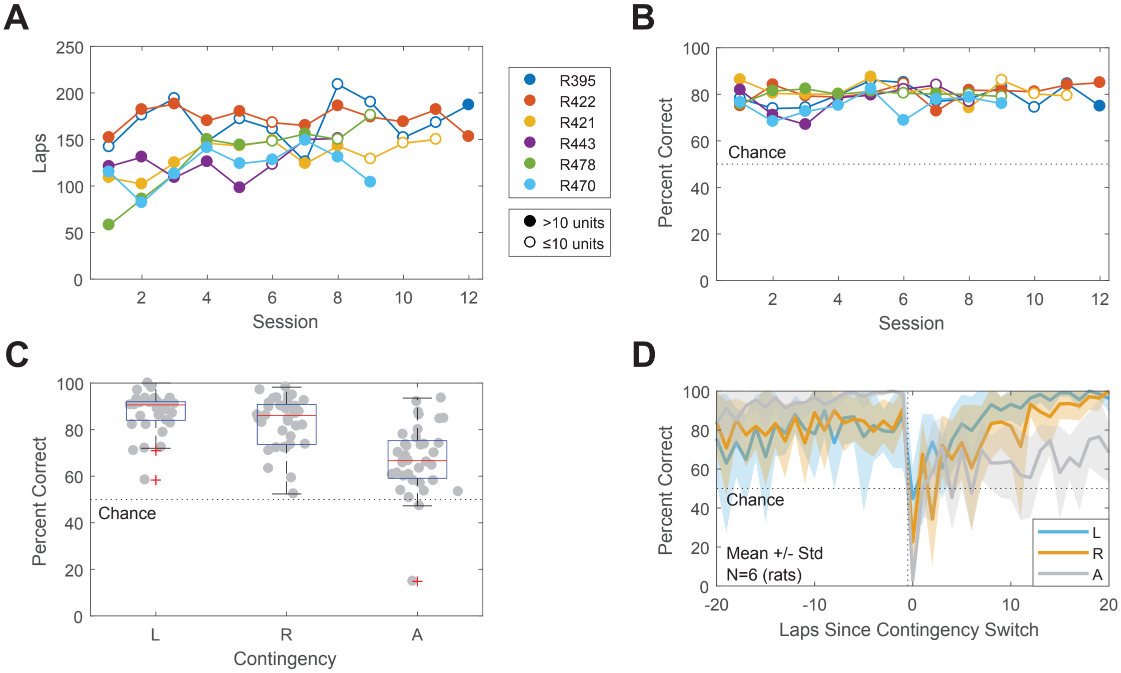

Figure 3.2: Behavioral performance on the MT-LRA task. (A) Laps per session ($N = 6$ rats). Filled circles indicate sessions which met the inclusion criteria (> 10 cells simultaneously recorded in both structures), and empty circles correspond to sessions which were not used for neural analyses. (B) Percent correct by session. (C) Performance by contingency ($N = 40$ sessions). (D) Performance aligned to switch, split by contingency.

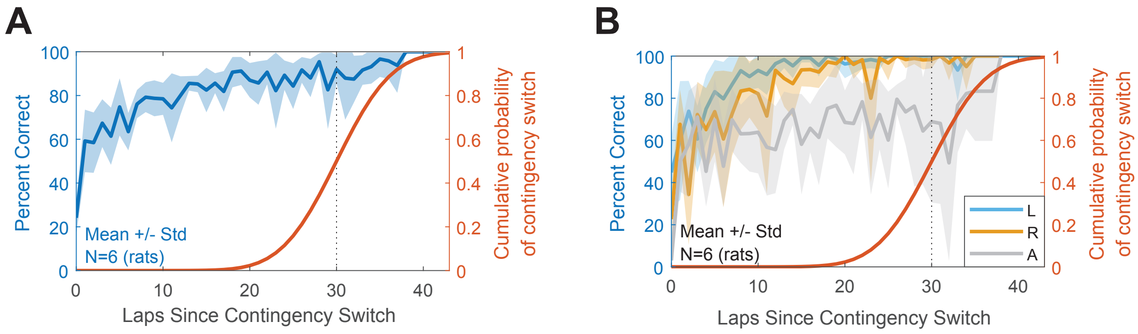

Figure 3.3: Performance aligned to the previous switch (A) overall, and (B) split by contingency. If rats were anticipating contingency switches, we would expect to see more probing of choices which were inconsistent with the current contingency, which would have led to a decrease in the proportion of correct trials as rats approached the expected transition lap.

any behavioral signs of anticipating the switch, as their choices did not reflect an increase in actions consistent with other contingencies as rats approached the expected contingency switch lap (Figure 3.3).

The percentage of correct choices dropped on laps immediately following a contingency switch, but then increased over the course of the following contingency block, and plateaued well before the next contingency switch (Figure 3.4A).

To identify laps where rats updated their behavioral choices to be consistent with the new contingency, we used a change point analysis from Gallistel et al. (2004). We considered 20 laps on either side of a contingency switch, after which the contingency in place was contingency X. We excluded laps which were before the previous switch, or after the next switch (in cases where contingency blocks lasted < 20 laps). For each lap i in this window around each switch, we computed whether rats' choices were consistent with the new contingency (the rat made a choice which would be correct if X were the current contingency, $c_i = 1$) or inconsistent with the new contingency (the rat made a choice which would be incorrect if X were the current contingency, $c_i = 0$). We then applied the change point analysis from Gallistel et al. (2004) on $\mathbf{c}$

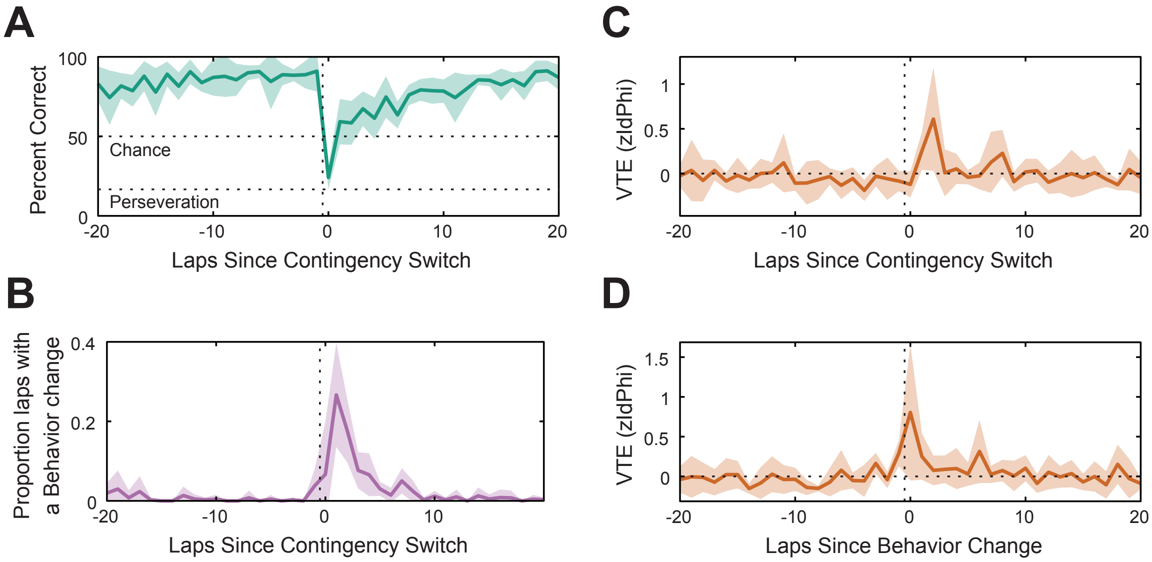

Figure 3.4: Rat behavior on the MT-LRA task aligned to contingency switches. (A) Rat performance aligned to contingency switches. The vertical dotted line corresponds to the last lap of the previous contingency block. (B) Rat behavioral change laps aligned to contingency switches. The dotted line corresponds to the last lap of the previous behavioral strategy. (C) VTE (measured by zIdPhi) aligned to contingency switches. (D) zIdPhi aligned to behavioral change laps. Plots show mean ± standard deviation, $N = 6$ rats.

to determine on what lap rats were most likely to have updated their choices to be consistent with the new strategy. This change-point analysis (Gallistel et al., 2004) indicated that rats updated their behavioral strategies to be consistent with the new contingency within about 5 laps of a contingency switch (Figure 3.4B).

While at choice points, rats sometimes display vicarious trial and error (VTE), a behavioral marker of deliberation (Redish, 2016), also see section 2.3.2 for more detail. During VTE behaviors, rats pause and look back and forth down potential paths, as if deliberating over which path to choose (Figure 3.5A). To quantify VTE, we measured *zIdPhi*, the z-scored integrated angular velocity of head movement (Papale et al., 2012). See section 2.3.2 for a more thorough definition of IdPhi.

To distinguish VTE events from non-VTE events, we fit a half-Gaussian distribution to values less than the mode of the zIdPhi distribution. We then assumed that zIdPhi values under a full Gaussian distribution with the same mean and standard deviation as the fit half-Gaussian corresponded to non-VTE events, and passes through the choice point with greater zIdPhi values corresponded to passes on which VTE occurred (Figure 3.5B).

As with many other studies which examine VTE (Steiner and Redish, 2012; Schmidt et al., 2013; Stott and Redish, 2014), we observed low levels of zIdPhi on most choice point passes and higher levels of zIdPhi on fewer laps (Figure 3.5B), suggesting rats deliberated on the minority of laps. A decrease in the amount of VTE over the course of a session is usually observed on other tasks (Papale et al., 2012; Breton et al., 2015; Redish, 2016), but on our task zIdPhi did not decrease over the course of the session (Figure 3.5C). This suggests that the presence of multiple contingency switches, which continued to occur throughout the course of the session, repeatedly forced rats to deliberate and prevented them from fully automating their behavior on the task.

Although rats did not appear to automate over the course of an entire session, they did automate over the course of single contingency blocks. On laps immediately following a contingency switch zIdPhi increased, and then decreased throughout the subsequent contingency block (Figure 3.4C). This suggests that rats deliberated after contingency switches, but then automated as they learned the new contingency. However, this effect seemed to be mostly driven by switches to the Alternate contingency (Figure 3.5D). The greatest levels of VTE were observed on laps where rats updated their behavioral strategies to be consistent with the new contingency (Figure 3.4D, the median zIdPhi was significantly greater on laps where a behavioral change occurred than on other laps, two-sided Wilcoxon rank sum test, $p = 2.7 \times 10^{-5}$, $N = 164$ behavior change laps vs 5344 non-change laps).

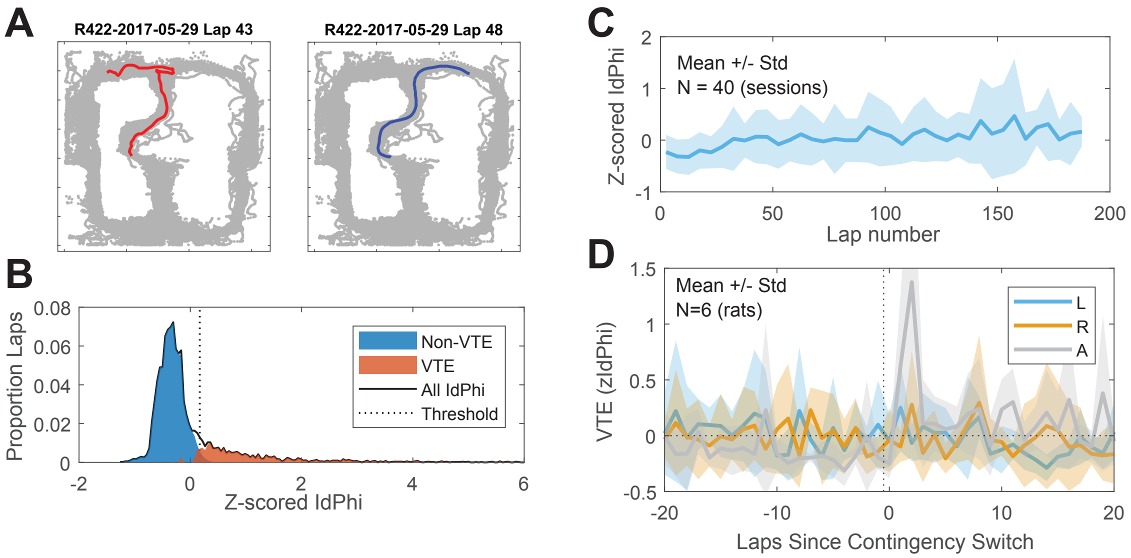

Figure 3.5: Vicarious trial and error (VTE) on the MT-LRA task. (A) Example of a pass through the choice point where the rat displayed VTE (left) and a non-VTE pass (right). (B) Distribution of zIdPhi across all laps and rats. (C) zIdPhi over the course of a session. (D) zIdPhi aligned to switch split by contingency.

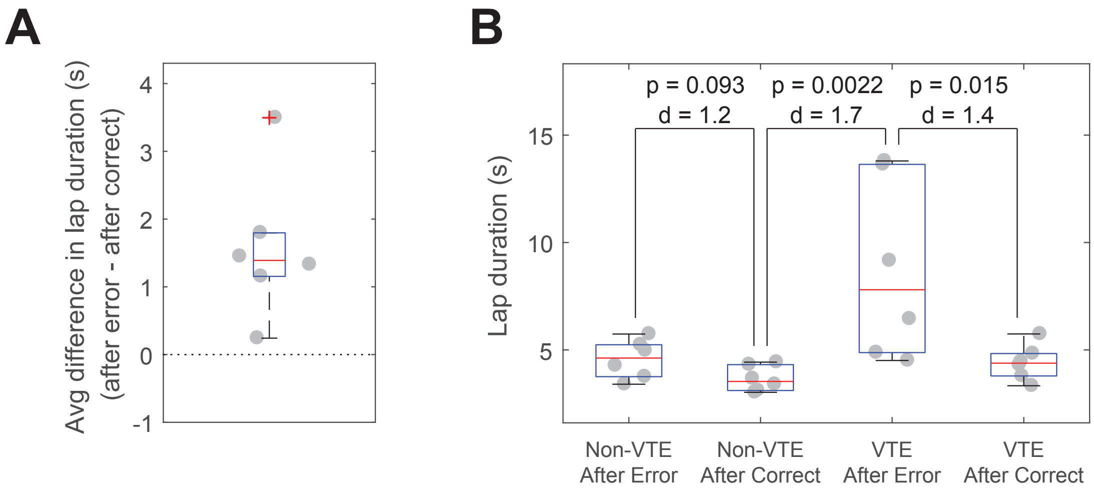

Figure 3.6: Post-error slowing. (A) The average difference in lap duration between laps following errors vs correct choices ($N = 6$ rats). (B) Lap duration split by both VTE and whether the rat made an error on the previous lap ($N = 6$ rats). Also shown are p values of two-sided Wilcoxon signed-rank tests, and Cohen's d.

Rats also displayed post-error slowing on our task (Laming, 1968; Narayanan and Laubach, 2008). Rats took significantly longer (around 1-2 seconds) to complete laps when the choice they made on the previous lap was incorrect (Figure 3.6A, two-sided Wilcoxon rank sum test, $p = 0.031$, $N = 6$ rats). Post-error slowing was especially pronounced on laps where VTE occurred (Figure 3.6B), which suggests that rats utilized a more conservative decision-making strategy following errors.

3.3 A Contingency-Aware Reinforcement Learning Algorithm

The fact that rats were proficient at performing the contingency switch task, and quickly adjusted their behavioral strategies to be consistent with new task contingencies within a few laps of contingency changes, indicated that the rats had some concept of the reward contingencies, and used information about those contingencies to guide their choices on the contingency switching task. In order to determine how rats kept track of contingency information, updated those beliefs, and used that information to make decisions, we designed and fit a reinforcement learning model to rat behavior on the contingency switch task.

In this section, we describe a contingency-aware reinforcement learning algorithm which formalizes a decision-making strategy which explicitly keeps track of the contingency probabilities, updates those beliefs after rewards (or reward omissions), and uses those contingency beliefs to make decisions at the choice point. We also fit this reinforcement learning algorithm to rat behavior in order to determine what algorithm parameters best explain the rats' choices. Like the reinforcement learning algorithms fit to behavior on the two step task, this contingency-aware algorithm computed the expected value (or Q-value) of taking an action a, in any given state, s. Our model of the contingency switch task included only two possible actions ("go left" or "go right"), and only one state: the choice point.

Some reinforcement learning models are able to adaptively cluster many different contexts into single contingency groups (Collins and Frank, 2013), and it is possible rats learned the contingencies via similar mechanisms. However, we built contingency knowledge into the algorithms under the assumption that if rats could learn the contingencies and use that information to inform their choices, then they would have learned the contingencies by the time the experiment started, since they were trained extensively on the contingency switching task before data collection began (see section 3.1 for details of the training schedule).

Instead of updating the Q-values for each potential action, the contingency-aware algorithm kept track of and updated the probability that each contingency was currently in place. That is, it stored P values for each contingency (the probability that that contingency was the current contingency), and updated those P values according to the agent's experience on the previous two laps.

On each lap immediately after reward delivery or lack thereof, the algorithm's internal probabilities of each contingency were updated by

$$P_c = P_c + \alpha\delta_c \tag{3.1}$$

where P_c is the algorithm's expectation of the probability that contingency c (L, R, or A) is currently in place. The α parameter is the learning rate, a free parameter.

For the left contingency, δ was computed by:

$$\delta_L = \begin{cases} 1 - P_L & \text{if } a_{i-1} = \text{left and } r_{i-1} = 1 \text{ and } a_{i-2} = \text{left and } r_{i-2} = 1 \\ 1 - P_L & \text{if } a_{i-1} = \text{left and } r_{i-1} = 1 \text{ and } a_{i-2} = \text{right and } r_{i-2} = 0 \\ 1 - P_L & \text{if } a_{i-1} = \text{right and } r_{i-1} = 0 \text{ and } a_{i-2} = \text{left and } r_{i-2} = 1 \\ 1 - P_L & \text{if } a_{i-1} = \text{right and } r_{i-1} = 0 \text{ and } a_{i-2} = \text{right and } r_{i-2} = 0 \\ -P_L & \text{otherwise} \end{cases} \tag{3.2}$$

where a_i is the action taken on trial i, and r_i is the reward experienced after taking an action on trial i.

Conversely, for the right contingency, δ was computed by:

$$\delta_R = \begin{cases} 1 - P_R & \text{if } a_{i-1} = \text{right and } r_{i-1} = 1 \text{ and } a_{i-2} = \text{right and } r_{i-2} = 1 \\ 1 - P_R & \text{if } a_{i-1} = \text{right and } r_{i-1} = 1 \text{ and } a_{i-2} = \text{left and } r_{i-2} = 0 \\ 1 - P_R & \text{if } a_{i-1} = \text{left and } r_{i-1} = 0 \text{ and } a_{i-2} = \text{right and } r_{i-2} = 1 \\ 1 - P_R & \text{if } a_{i-1} = \text{left and } r_{i-1} = 0 \text{ and } a_{i-2} = \text{left and } r_{i-2} = 0 \\ -P_R & \text{otherwise} \end{cases} \tag{3.3}$$

For the alternate contingency, δ was computed by:

$$\delta_A = \begin{cases} 1 - P_A & \text{if } a_{i-1} \neq a_{i-2} = \text{ and } r_{i-1} = 1 \text{ and } r_{i-2} = 1 \\ 1 - P_A & \text{if } a_{i-1} = a_{i-2} = \text{ and } r_{i-1} = 1 \text{ and } r_{i-2} = 0 \\ 1 - P_A & \text{if } a_{i-1} \neq a_{i-2} = \text{ and } r_{i-1} = 0 \text{ and } r_{i-2} = 1 \\ -P_A & \text{otherwise} \end{cases} \tag{3.4}$$

The contingency-aware algorithm also contained a "forgetting" parameter (ϕ), which decayed the P-values to the baseline value each lap:

$$\forall c,\ P_c = P_c + \phi(\frac{1}{3} - P_c) \tag{3.5}$$

After P-value updating, these P values were transformed into a proper probability distribution across contingencies via a softmax with temperature parameter β_c:

$$P_c = \frac{\beta_c P_c}{\sum_{c' \in \{L,R,A\}} \beta_{c'} P_{c'}} \tag{3.6}$$

Then, the Q-value for each action was computed from these contingency probabilities. For the left action, the Q value was updated with:

$$Q(\text{left}) = \begin{cases} P_L + P_A & \text{if } a_{i-1} = \text{right} \\ P_L & \text{otherwise} \end{cases} \tag{3.7}$$

The right action's Q value was updated with:

$$Q(\text{right}) = \begin{cases} P_R + P_A & \text{if } a_{i-1} = \text{left} \\ P_R & \text{otherwise} \end{cases} \tag{3.8}$$

To transform the algorithm's valuations of different actions (the Q-values) into probabilities that the algorithm would make the same choice as the rats did on trial i (we denote this probability by $p(a_i)$), we used a softmax over the Q-values:

$$p(a_i) = \frac{\exp(\beta_v[Q(a_i) + p \times rep(a_i) + b \times type(a_i)])}{\sum_{a'} \exp(\beta_v[Q(a') + p \times rep(a') + b \times type(a')])} \tag{3.9}$$

where β_v is an inverse temperature parameter that controls how stochastic the models' choices are at each choice point, and the sum in the denominator sums over all available actions, a' (in the case of the contingency switch task, just left or right).

The p parameter accounts for an inclination to repeat the same action taken on the last lap ($p > 0$), or to switch to the opposite action ($p < 0$), regardless of expected action values. $rep(a')$ was a function which evaluated to 1 if the rat repeated its action, that is, performed action a' on the previous lap (trial $i - 1$), and 0 if it chose a different action. Therefore if the p parameter was positive, the algorithm was more likely to repeat the previous choice, and if it was negative, the algorithm was more likely to switch (choose the opposite choice from the previous trial). The purpose of this p parameter was to capture perseveration behavior.

Also, the b parameter accounts for side biases. The $type(a)$ term evaluated to 1 if action a was left:

$$type(a) = \begin{cases} 1 & \text{if } a = \text{left} \\ \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

Therefore, if the b parameter was negative, agents preferred choosing left, and if it was positive, they preferred choosing right. If it was 0, there was no side bias.

We initialized all P values to $\frac{1}{3}$. The log likelihood of observing rat choices across all N_s sessions given the algorithm and parameter values was then computed by summing the log likelihood of each choice for each lap, session, and rat:

$$\log(p(\text{data}|\theta)) = \sum_{k=1}^{N_r} \sum_{j=1}^{N_s} \sum_{i=1}^{N_l} \log\left(p(a_i)\right)) \tag{3.11}$$

where θ is the set of all parameters, N_l is the number of laps in a session, N_s is the number of sessions for a rat, and N_r is the number of rats.

To determine what algorithm parameter values best explained rats' choices, we performed Bayesian inference using Markov chain Monte Carlo (MCMC) in Stan

Parameter	2.5%	50%	97.5%	$\hat{R}$
α	0.533	0.623	0.700	1.000
β_c	5.41	7.17	10.8	1.000
β_v	1.62	1.67	1.72	1.000
p	-0.0202	-0.00766	0.00487	1.000
b	0.422	0.448	0.475	1.000
ϕ	0.00491	0.0752	0.255	1.000

Table 3.2: Contingency-aware algorithm fit to rat behavior. Percent columns indicate the bottom (2.5%), middle (50%), and top (97.5%) of the inner 95% posterior credible interval. The $\hat{R}$ statistic measures MCMC chain convergence, and should be between around 0.9 and 1.1 if MCMC chains have successfully converged.

(Carpenter et al., 2017). We used the Python programming language interface to Stan, PyStan (Stan Development Team, 2017), to generate model parameter posterior distributions so that we could perform inference as to the parameter values (Kruschke, 2014). The results of the fit are shown in Table 3.2.

3.4 VTE is likely related to Contingency Uncertainty

How did vicarious trial and error correspond to the rats' uncertainty as to the identity of the current contingency? To estimate rats' contingency uncertainty, we simulated the contingency-aware algorithm (see section 3.3) using the maximum a posteriori parameter values from the fits to rat behavior. The simulations were presented with the same choice and reward sequences as the rats, and from the algorithms we computed contingency uncertainty on each lap i as the entropy of the contingency probabilities:

$$u_i = H(\{P_L(i), P_R(i), P_A(i)\}) \tag{3.12}$$

where H denotes the information entropy over the discrete set of contingencies:

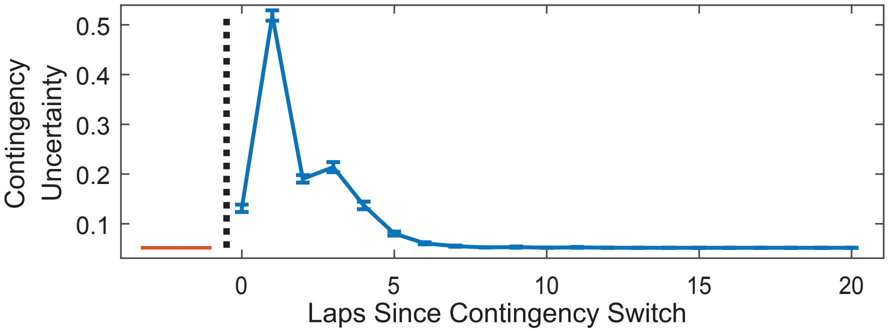

Figure 3.7: Contingency uncertainty aligned to the switch

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{3.13}$$

Similarly to vicarious trial and error, the contingency uncertainty increased after contingency switches and then decreased over the course of contingency blocks (Figure 3.7).

But were vicarious trial and error and contingency uncertainty truly correlated, or was the similarity in their timecourses due simply to the recency of a contingency change, or the time within a session, or other factors?

To address this question, we fit a multilevel model of vicarious trial and error at the choice point of the contingency switch task. In this model, LogIdPhi (a measure of vicarious trial and error, see section 2.3.2) was predicted by trial within a session, the recency of a contingency switch, the contingency uncertainty (equation 3.12), but also included a random effect of rat.

$$Y_i \sim \mathcal{N}(\beta_0 + R_r + T_r t_i + S_r s_i + U_r u_i,\ \sigma_e) \tag{3.14}$$

where R_r is a random effects coefficient for rat r, and T_r, S_r, and U_r are per-rat parameters drawn from population distributions corresponding to random effects of trial, switch recency, and uncertainty, respectively.

$$\begin{aligned} R &\sim \mathcal{N}(0, \sigma_r) \\ T_r &\sim \mathcal{N}(\mu_t, \sigma_t) \\ S_r &\sim \mathcal{N}(\mu_s, \sigma_s) \\ U_r &\sim \mathcal{N}(\mu_u, \sigma_u) \end{aligned} \tag{3.15}$$

where

- Y_i is the LogIdPhi value at the choice point on lap i,
- β_0 is the intercept of the model (baseline LogIdPhi value),
- μ_t is the mean effect of trial on LogIdPhi,
- t_i is the trial number on lap i,
- μ_s is the mean effect of contingency switch recency on LogIdPhi,
- s_i is the number of laps since the last contingency switch on lap i,
- μ_u is the mean effect of contingency uncertainty on LogIdPhi,
- u_i is the contingency uncertainty on lap i (equation 3.12),
- σ_e is the standard deviation of the error, and
- $\mathcal{N}(\mu, \sigma)$ represents a normal distribution centered at μ with standard deviation σ.

We found that uncertainty was likely related to vicarious trial and error at the choice point of the contingency switch task. Although the center 95% credible interval of the posterior distribution for the population mean of the uncertainty effect on

$$\text{LogIdPhi}_i \quad = \quad \beta_0 + R_r + T_r t_i + S_r s_i + U_r u_i + \epsilon$$

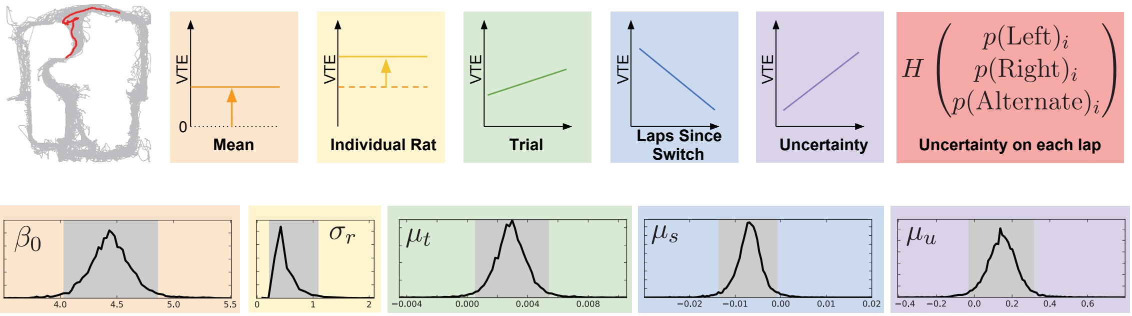

Figure 3.8: The mixed model of VTE and uncertainty. Shaded areas in the lower panels show the middle 95% credible intervals of the posterior distributions for each parameter of interest.

LogIdPhi included 0, 96.3% of the posterior density was greater than 0 (Figure 3.8). This suggests that the amount of VTE displayed by the rats was very likely related to their uncertainty as to the current contingency.

All this suggests that rats indeed were keeping track of the task contingencies and using that information to make contingency-informed decisions on the switch task. But where in the brain was this information stored, and how was it retrieved during deliberation? In the next section, we look at contingency representation in the hippocampus and prefrontal cortex as rats run the contingency switch task, and determine which aspects of neural activity can be explained by explicit representation of the contingency identities, and which other aspects are simply the result of unrelated representational changes over time.

Chapter 4

CA1 and dmPFC Encode Both Contingencies and Time

4.1 Introduction

The deliberative system is thought to involve many different brain areas which collectively represent abstract information about the environment, and which interact to use that information to inform decision making during dynamic or difficult action selection challenges. Along with other structures, the hippocampus (HPC) and the prefrontal cortex (PFC) represent spatial information, information about more abstract task contingencies, and other information which may be changing over time (such as information related to motivational state, arousal, hunger, etc). But what are the dynamics of the flow of information between these two structures, and how can the representations of abstract contingencies be disentangled from other information representation? In this chapter we investigate contingency representations in the dorsomedial prefrontal cortex (dmPFC) and the first subfield of the cornu Ammonis of the hippocampus (CA1), how representations in these areas change over time, and how encoding due to these two factors can be parsed apart.

The prefrontal cortex as long been thought to mediate executive function (Miller and Cohen, 2001; Dalley et al., 2004; Kesner and Churchwell, 2011). It participates in the storage and recall of contextual memories (Tronel and Sara, 2003; Euston et al., 2012; Preston and Eichenbaum, 2013) and maintains that information in working memory (Ragozzino and Kesner, 1998; Delatour and Gisquest-Verrier, 1999; Cowen and McNaughton, 2007; Yoon et al., 2008; Horst and Laubach, 2009; Urban et al., 2014).

Specifically, dorsomedial aspects of the prefrontal cortex (dmPFC) play several different roles which support behavioral flexibility. In rodents, the dmPFC is comprised of three main subregions. From most dorsal to most ventral they are: the anterior cingulate cortex (ACC), prelimbic cortex (PL), and infralimbic cortex (IL). The dmPFC is important for conflict resolution, and especially when the conflicts involve abstract or latent factors, as lesioning the dmPFC interferes with animals' abilities to detect and inhibit inappropriate responses (Haddon and Killcross, 2005, 2006). ACC is traditionally associated with behavioral inhibition, though this responsibility appears to be somewhat distributed between subregions, as inactivation of PL also impairs the control of contextually-dependent behaviors (Marquis et al., 2007; Dwyer et al., 2010).

The dmPFC is also required for learning task contingencies and adjusting behavioral strategies accordingly. Neurons in dmPFC encode abstract task rules (Balleine and Dickinson, 1998; Jung et al., 1998; Wallis et al., 2001; Hyman et al., 2012), and also appear to represent information about context (Mante et al., 2013; Powell and Redish, 2014; Ma et al., 2016). Also, in rodents, inactivating dmPFC prevents animals from updating their behavioral strategies to match changing task rules (Ragozzino et al., 2003; Floresco et al., 2008; Young and Shapiro, 2009).

Neural activity in dmPFC reflects changes in behavioral strategy (Rich and Shapiro, 2009; Karlsson et al., 2012; Powell and Redish, 2016; De Falco et al., 2019). However, it is unclear whether the dmPFC only encodes changes in behavioral strategies or task contingencies, as opposed to actually carrying information about the identity of the task rules (Durstewitz et al., 2010; Malagon-Vina et al., 2018).

Also, the dmPFC contributes to decision-making and generating goal-directed actions. However, there are subregion-specific differences in these contributions to goal-directed decision making (Seamans et al., 1995). Specifically the prelimbic and infralimbic subregions have been found to be important for goal-directed behaviors, and represent goal-relevant information (Matsumoto et al., 2003; Matsumoto and Tanaka, 2004; Hok et al., 2005; St. Onge and Floresco, 2009). These areas may even contribute to the balancing of habitual and deliberative influences on action selection (Killcross and Coutureau, 2003).

Algorithmically, how might the dmPFC contribute to deliberation? Theoretical work supported by some experimental work suggests that the dmPFC may control internal simulations of possible actions and their outcomes, and use evaluations of the internally simulated outcomes to inform action selection (Hassabis and Maguire, 2009; van der Meer et al., 2012; Wang et al., 2015).

On the other hand, while hippocampus (HPC) is traditionally thought to represent spatial location and to play a central role in spatial navigation, the activity of neurons in HPC also reflect cognitive, non-spatial information. The hippocampus has long been known to be important for either storing or recalling episodic memories (Scoville and Milner, 1957; O'Keefe and Nadel, 1978b; Cohen and Eichenbaum, 1993; Redish, 1999). In rodents, hippocampal representations have primarily been studied in the context of "place cells".

Hippocampal place cells are cells in the CA1 and CA3 subregions of the hippocampus which have high spatial selectivity (O'Keefe and Dostrovsky, 1971; O'Keefe and Nadel, 1978b). That is, their tuning curves are tightly tuned to specific spatial locations: the cells spike at high rates when rodents are in a specific location, and are nearly silent when the animal is in other locations. The size of hippocampal place fields vary greatly, and differ depending on the environment, task, and context, but in the dorsal hippocampus of rodents fields generally range in size between a few centimeters to a meter.

However, the hippocampus also encodes other more abstract information which is not purely spatial. For example, hippocampal place fields in rats depend on where the rat intends to go – on spatial alternation tasks where rats run to one side of a figure-eight maze on one lap, and then to the opposite side on the next lap, place field activity on the central arm displays "splitter" behavior, where firing of the place cell in its field is dependent on the side the rat is about to run to (Wood et al., 2000; Ferbinteanu and Shapiro, 2003; Smith and Mizumori, 2006). Place fields are also modulated by context (Hasselmo and Eichenbaum, 2005; Griffin et al., 2007; Zilli and Hasselmo, 2008; Kennedy and Shapiro, 2009; Ferbinteanu et al., 2011). Furthermore, place cells sometimes even completely remap (randomly change their firing field locations) or modulate the maximum firing rates of their fields depending on sensory cues (Sharp et al., 1990, 1995; Leutgeb et al., 2005; Bahar et al., 2011).

Hippocampus and dmPFC, along with other structures, are thought to form a processing loop where bottom-up information from HPC informs representations in dmPFC, and top-down signals from dmPFC influence memory retrieval in HPC based on context or strategy (Wang et al., 2015; Jai and Frank, 2015; Redish, 2016; Shin and Jadhav, 2016; Eichenbaum, 2017).

How do the hippocampus and the dmPFC communicate? There are both direct and indirect projections from HPC to dmPFC (Swanson, 1981; Ferino et al., 1987; Jay and Witter, 1991; Verwer et al., 1997; Delatour and Witter, 2002; Floresco and Grace, 2003; Hoover and Vertes, 2007), as well as bidirectional connections between dmPFC and HPC via the nucleus reuniens of the thalamus and other thalamic nuclei (Vertes, 2002, 2004; McKenna and Vertes, 2004; Vertes et al., 2006; Di Prisco and Vertes, 2006; Vertes et al., 2007; Hoover and Vertes, 2012; Cassel et al., 2013; Dolleman-Van der Weel et al., 2017; Dolleman-van der Weel et al., 2019).

Contralateral lesion studies suggest that a PFC-HPC connection is required for spatial working memory. These studies inhibited the PFC in one hemisphere, and the hippocampus in the opposite hemisphere, and found that animals with these contralateral "lesions" showed performance deficits on tasks requiring spatial working memory (Floresco et al., 1997; Wang and Cai, 2006, 2008; Barker et al., 2017; Maharjan et al., 2018). While these experiments are not able to perfectly simulate the disruption of inter-area communication (HPC could project indirectly to the contralateral dmPFC and vice-versa), the suggest that information transfer between these two areas is important for maintaining and using spatial memories.

Electrophysiology experiments also support theories that there are interactions between HPC and dmPFC during working memory tasks. Local field potential oscillations in the dmPFC and hippocampus become synchronized during decision-making requiring working memory, in the theta (6-11Hz), gamma (40-80Hz), and even perhaps delta (¡4Hz) frequency bands (Siapas et al., 2005; Jones and Wilson, 2005b; Sirota et al., 2008; Hyman et al., 2010; Colgin, 2011; Gordon, 2011; O'Neill et al., 2013; Fujisawa and Buzsáki, 2011; Brincat and Miller, 2015; Place et al., 2016; Liu et al., 2018), and dmPFC or HPC oscillations lead or lag each other depending on whether the animal is recalling information at decision time or encoding information during exploration (Place et al., 2016; Liu et al., 2018). Spike timing in PFC also appears to synchronize with hippocampal theta rhythms, again especially at times when memory recall is required during decision making tasks (Hyman et al., 2005; Jones and Wilson, 2005a; Sirota et al., 2008; Benchenane et al., 2010; Hyman et al., 2011; Spellman et al., 2015; Zielinski et al., 2019).

Disrupting activity in PFC changes information representation in HPC (Hok et al., 2013; Navawongse and Eichenbaum, 2013; Guise and Shapiro, 2017; Schmidt et al., 2019). This effect is thought to be mediated by the nucleus reuniens (Dolleman-van der Weel et al., 2009; Hallock et al., 2013; Xu and Südhof, 2013; Griffin, 2015; Ito et al., 2015; Layfield et al., 2015; Linley et al., 2016; Hallock et al., 2016; Ito et al., 2018; Viena et al., 2018; Mei et al., 2018; Maisson et al., 2018; Zimmerman and Grace, 2018). Conversely, disrupting HPC outputs to the PFC weakens the encoding of spatial working memories (Spellman et al., 2015).

How does information about environmental contingencies flow between prefrontal cortex and hippocampus? Specifically, do representations in dmPFC reflect new contingencies before, after, or at the same time as representations in HPC? Current theories which posit that dmPFC exerts a top-down contextual influence on HPC certainly lead to the hypothesis that updates in contingency or strategy representations would be seen in dmPFC before HPC. But even if representations of new rules in dmPFC do precede those in HPC, the time scale of this lead is difficult to deduce: it could be anywhere from tens of millisecond to minutes. Guise and Shapiro (2017) examined the interaction between contingency representations by HPC and mPFC ensembles on a task with rule switches. They provided convincing evidence that both HPC and mPFC represent goals, and that some HPC activity is predictable from past mPFC activity on trials after rule changes. However, the study did not investigate the time course or latency of this interaction, nor did it identify specifically what information in mPFC facilitated the prediction of HPC activity.

In addition to representing contingency or strategy information, ensemble activity in both HPC and PFC has been found to change slowly over time (Mankin et al., 2012; Hyman et al., 2012; Ziv et al., 2013; Malagon-Vina et al., 2018). Some theories suggest that this change over time provides a form of temporal context (Mensink and Raaijmakers, 1988; Howard and Kahana, 2002), because events which occur close in time would have similar representations or “timestamps” (Rubin et al., 2015), leading to easier retrieval of temporally similar memories. The change of ensemble encoding over time could also be due to the representation of other unmeasured factors such as motivation or satiation. Alternatively, the change could simply be due to a random drift in representations over time. While this representational drift appears to occur in both HPC and dmPFC, it is unknown how quickly representations drift in each structure relative to the other. Furthermore, drifting neural activity over time could be misconstrued as task rule representation when the rules are presented in blocks of time, and so it remains unclear how much of contingency or strategy representations in block-structured tasks can be explained simply by representational drift.

In this chapter, we recorded from neural ensembles in dmPFC and dorsal CA1 simultaneously as rats performed the contingency switching task, in order to investigate how the two areas represent task contingencies, information which is hypothesized to only be used by the model-based system. We developed an analysis to disambiguate the contributions of contingency representation from the representation of other non-contingency time-varying information. We determined that representations in dmPFC and HPC encode task contingencies while simultaneously changing over time in ways unrelated to contingency, and that contingency representation could not be explained by this encoding drift. We also compared the time course of contingency encoding changes between dmPFC and HPC as task rules changed, and compared the rate of change of representations in both structures.

4.2 Surgery and Neural Recordings

4.2.1 Surgery and targeting

After training on the contingency switch task (see section 3.1), rats were given free access to food for at least 3 days, and then chronically implanted with a hyperdrive containing 24 tetrodes (built in-house), and a separate drive containing a 32-site silicon probe (Cambridge NeuroTech, Cambridge, England). The hyperdrives contained two bundles of 12 tetrodes each, targeting the CA1 region of dorsal hippocampus bilaterally (3.8 mm posterior and ± 3.0 mm lateral from bregma). The hyperdrive for one rat contained a single bundle of 24 tetrodes targeting the right hippocampus. The silicon probes consisted of two 16-site shanks which were implanted 3.8 mm anterior to and 0.7 mm lateral from bregma at a 25 degree angle (targeting dmPFC on the right hemisphere, such that the final target was 2.3 mm A/P, -0.7 mm M/L, and 3.9mm D/V, all coordinates relative to bregma). The hyperdrives and silicon probe drives were made in-house, and protective shrouds around the drives and amplifier boards were printed on a Form 2 3D printer (Formlabs, Somerville, MA).

Animals were anaesthetized with and maintained on isoflurane ($0.5 - 2\%$ isoflurane vaporized in O_2) for the duration of the surgery. Rats were placed in a sterotaxic apparatus (Kopf, Tujunga, CA) and were given penicillin (Combi-Pen-48) intramuscularly in each hindlimb, and carprofen (Rimadyl) subcutaneously. Rats' heads were shaved and disinfected with Betadine (Purdue Rederick, Norwalk, CT) before making an incision to reveal the skull. 3-5 jewlers' screws were used to anchor the drives to the skull, one of which was used as ground for the tetrodes, and a separate screw used as ground for the probe.

Three craniotomies were opened: two for the bilateral tetrode bundles using a surgical trephine, and one for the silicon probes using a burr. The dura was removed with forceps, the probe and tetrode drives were positioned with the sterotax, and then silicone gel (Dow Corning, Midland, MI) was applied to the craniotomies. A layer of MetaBond (Parkell, Edgewood, NY) and then dental acrylic (The Hygenic Corporation, Cuyahoga, OH) was applied to secure the drives to the skull. After surgery, the probes and tetrodes were turned down 640 μm. Rats were subcutaneously injected with carprofen on the day of surgery and for 2 days after surgery, as well as enrofloxacin (Enroflox) the day of surgery and for 5 days post-surgery.

4.2.2 Data acquisition and electrophysiology

Neural data from all rats was acquired on an Intan RHD2000 recording system (Intan Technologies, Los Angeles, CA), using four RHD2132 amplifier boards (three for the tetrodes and one for the silicon probe). The digitized signals were passed through a 24-channel commutator (Moog, East Aurora, NY) to allow the rats to move freely throughout recording sessions. To synchronize behavior with the neural recordings, timestamps were sent from the Matlab (Version 2017a, The MathWorks, Inc., Natick, MA) software running the task to digital input ports on the Intan RHD2000 USB Interface Board via an Arduino Uno.

Tetrodes were slowly advanced toward the hippocampal pyramidal layer, and the probes toward dmPFC, over the course of around 2 weeks, as the rats recovered and were re-trained on the task. The pyramidal layer was identified by the size of ripples and the direction of sharp wave deflection, as well as spike bursts during these SWRs.

Signals were filtered and spikes and LFP signals were extracted using in-house software written in Matlab and C. Spikes recorded on tetrodes in the hippocampus were manually clustered using the MClust 4.4 software package (Redish, 2017). Only well-separated clusters were kept and used for analysis. The median isolation distance

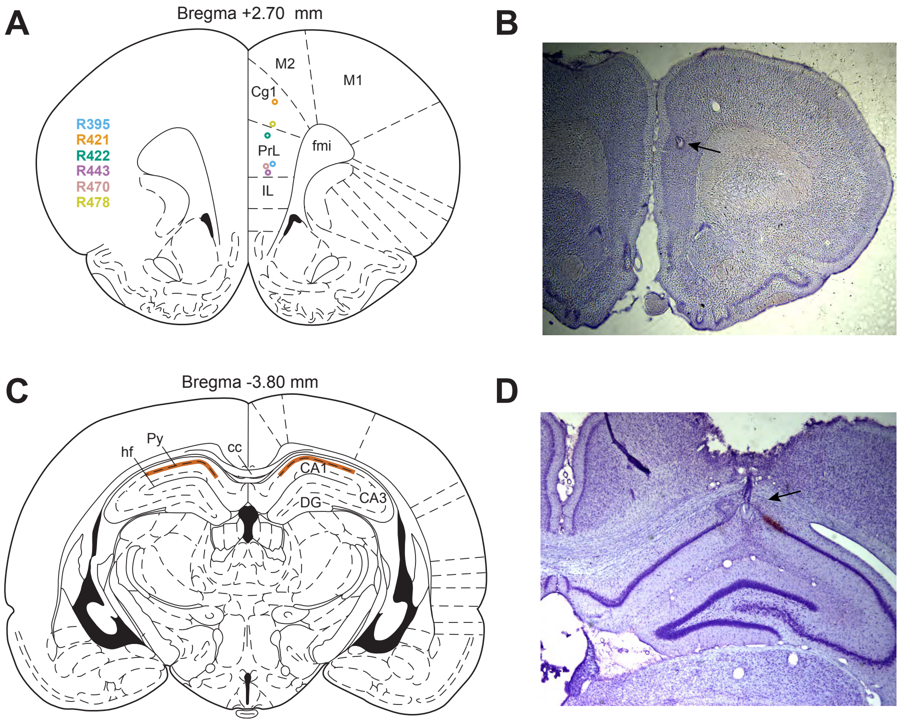

Figure 4.1: Histology and electrode targeting. (A) Si probe recording locations in dmPFC for each rat. (B) Example photo of a cresyl violet stained coronal slice through PFC, showing the electrolytic lesion created to mark the recording location (in Cg1 in this example). (C) Tetrode recording region (highlighted) was in the pyramidal layer of dorsal CA1. (D) Example photo of a cresyl violet stained coronal slice through HPC, showing an electrolytic lesion and electrode track. Anatomy diagrams in (A) and (C) are from Paxinos and Watson (2006). *IL*, infralimbic cortex. *PrL*, prelimbic cortex. *Cg1*, cingulate cortex area 1. *M2*, secondary motor cortex. *M1*, primary motor cortex. *fmi*, forceps minor of the corpus callosum. *hf*, hippocampal fissure. *Py*, pyramidal layer. *cc*, corpus callosum. *DG*, dentate gyrus.

was 21.2, and the median L-Ratio was 0.0899 (Schmitzer-Torbert et al., 2005). Spikes recorded on silicon probes were sorted offline using Kilosort (Pachitariu et al., 2016) into putative clusters, and then manually refined using Phy (Rossant et al., 2016).

4.2.3 Histology

After rats were finished running the experiment, both tetrode and silicon probe recording locations were marked with electrolytic lesions. 10μA was passed through a chan-

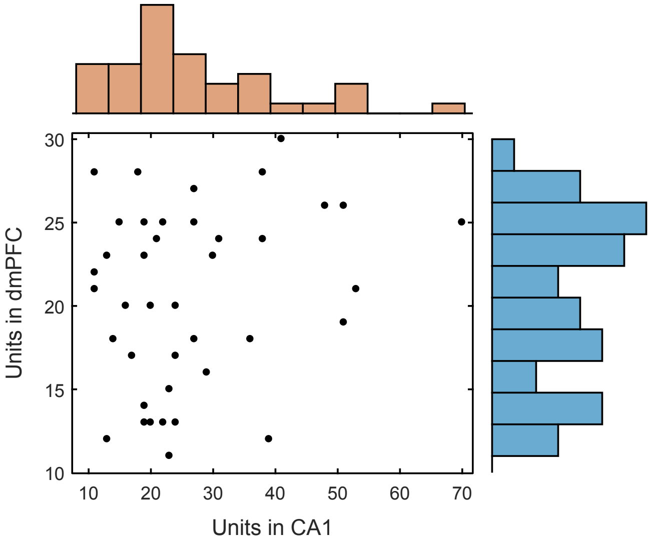

Figure 4.2: Number of cells recorded simultaneously per session in dmPFC and in CA1.

nel on each tetrode, and every fourth channel on the silicon probes, for 10s. At least two days after the lesions were made, the rats were anesthetized with a pentobarbital sodium solution (150 mg/kg, Fatal-Plus) and then perfused transcardially with saline followed by 10% formalin. Brains were stored in formalin, and then in a 30% sucrose formalin solution until slicing. Coronal slices were made through the hippocampus and prefrontal cortex (sagittal slices were made instead in PFC for 4 rats) using a cryostat, and the slices were stained with cresyl violet and imaged to determine tetrode and silicon probe recording locations (Figure 4.1).

4.3 Ensemble activity represented both contingency and temporal context

We recorded simultaneously from neural ensembles in CA1 and dmPFC using two 16-site silicon probes (in dmPFC) and 24 tetrodes (in CA1) per animal. For all neural analyses, we included only sessions where ≥ 10 cells were recorded in dmPFC *and* ≥ 10 cells were recorded in CA1 simultaneously. During each of 40 sessions, we recorded 20.6 ± 5.4 units in dmPFC and 26.9 ± 13.5 units in CA1 (Figure 4.2).

To determine the extent to which ensemble representations reflected contingency encoding versus the encoding of other information which changed over time, we took advantage of the fact that our task contained multiple contingency blocks of the same type within a session. We analyzed the stability of contingency representations across multiple presentations of the same contingency. Specifically, we examined ensemble activity during the first presentation of a contingency of a given type (contingency block Y_1, Figure 4.3A), during a second presentation of that same contingency (contingency block Y_2), and during contingency blocks between the two (contingency block(s) X, during which the contingencies were of a different type). If time-varying information not related to contingency identity dominated the representations, ensemble activity during Y_1 and Y_2 should have been more dissimilar to each other than to ensemble activity during X, because they were further apart in time. On the other hand, if the contingency identities were encoded and other time-varying information, did not dominate the representations, then ensemble activity during Y_1 and Y_2 should have been more similar to each other than to ensemble activity during X. Our dataset contained $N = 62$ of these contingency epoch triplets, including a total of 4128 laps.

We used 20-fold cross-validated linear discriminant analysis (LDA) to project ensemble firing rates in dmPFC and CA1 onto the axis which best discriminated neural activity during Y_1 from that observed during X (Figure 4.3A, on the horizontal axis), and also onto the axis which best discriminated neural activity during Y_2 from that observed during X (Figure 4.3A, on the vertical axis). Projections were normalized on both axes such that -1 corresponded to the mean of projections during the Y epoch, and $+1$ corresponded to the mean of the projections during the X epoch. We

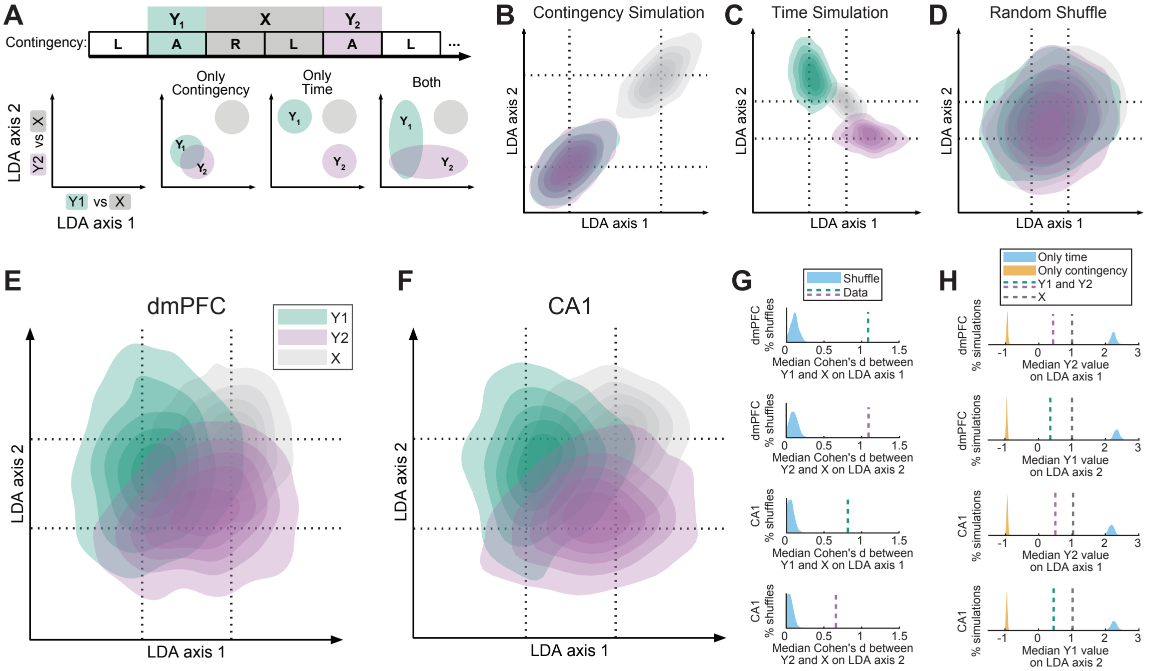

Figure 4.3: Analysis of the stability of contingency representations. (A) Illustration of the representational stability analysis. (B-F) show kernel-density-smoothed contour density plots of the LDA projections of neural activity on each lap of each contingency pair used in the analysis (contingency pairs of the same type separated by block(s) of another type). $N = 62$ contingency epoch triplets, including a total of 4128 laps. Example simulations and shuffles in (B-D) used a matched number of contingency epoch triplets, laps, and units. (B) LDA projections of simulated firing rates containing only contingency information. (C) LDA projections of simulated firing rates containing only time information. (D) LDA projections of randomly shuffled neural activity in dmPFC. (E) LDA projections of neural activity in dmPFC, and (F) CA1. (G) Cohen's d between projections during X epochs and Y_1 epochs (green dotted bars) or Y_2 epochs (purple dotted bars), as compared to projections of shuffled neural activity (blue distributions). (H) Median LDA projection values for neural activity during Y_1 (green dotted bar), Y_2 (purple dotted bar), or X (gray dotted bar), as compared to simulations where firing rates contained only contingency information (orange distributions), or simulations where firing rates were controlled only by temporal drift (blue distributions). Dotted lines in (G) and (H) correspond to the median projection across the whole dataset, and each sample making up the distributions corresponds to the median of a shuffle or simulation with a matched number of cells, laps, and contingency triplets.

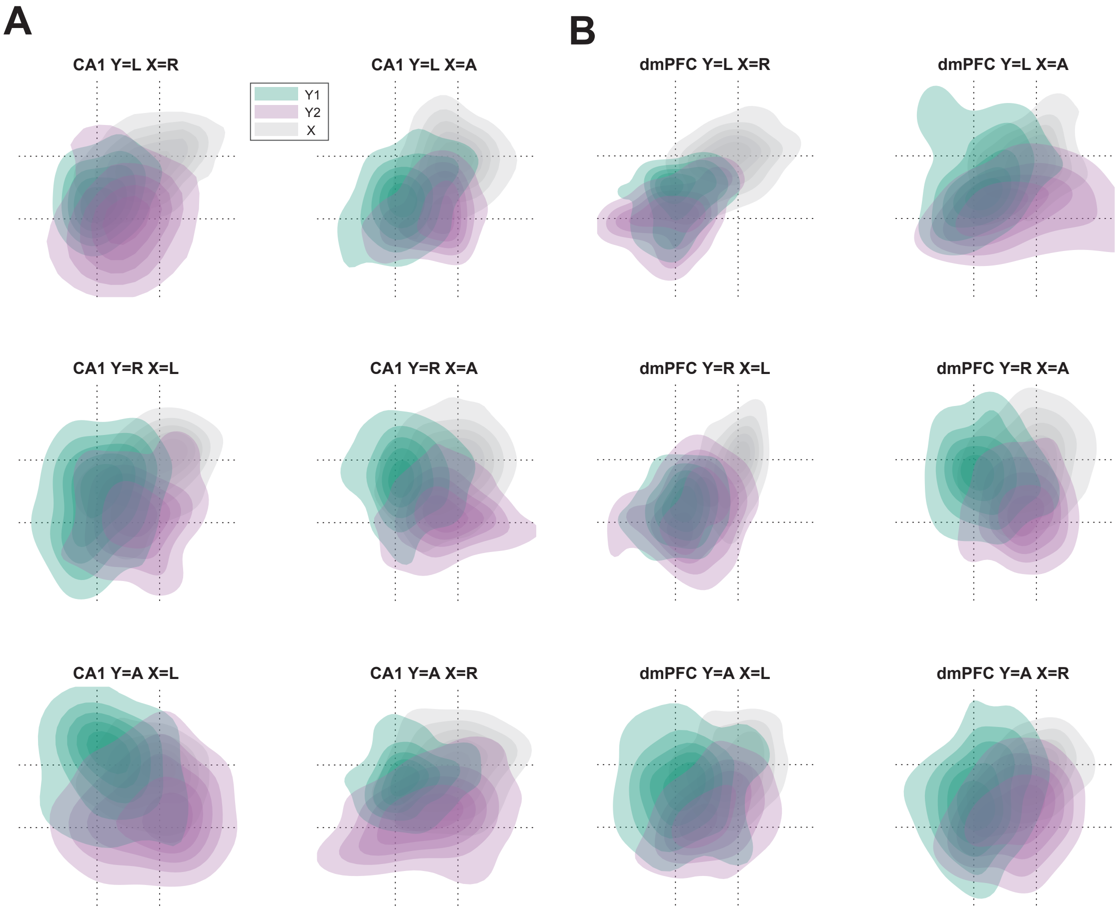

Figure 4.4: LDA projections in (A) CA1 and (B) dmPFC (as in Figure 4.3) for each transition type separately, for contingency triplets with only one intervening block. Y indicates the contingency type of the first and recurring contingency block, and X indicates the intervening block type.

included only the last 20 trials of each contingency block, with the intent that this included laps only after rats had learned the true contingency. This is similar to what was done in Malagon-Vina et al. (2018).

This is not to say that we believe contingency representations and other representations which change over time were combined solely linearly in both brain areas. We used LDA not as a statistical tool, but to provide a metric of the separation of ensemble firing rates between contingency blocks.

To validate that this analysis was able to disambiguate contingency information from other time-varying information, we first applied the analysis to simulated data where firing rates represented only the contingency, and were otherwise time-invariant. To generate firing rate simulations which represented only contingency and not time, we randomly assigned each cell a firing rate for each contingency, and added a small amount of noise. Again, this ensured that we generated "simulations" which had the same number of cells and sessions as our actual data, as well as identical firing rate distributions as our actual data, but had firing rates which represented only the current contingency. Projections of these simulated firing rates during Y_1 and Y_2 overlapped, but were well-separated from projections of simulated firing rates during X (Figure 4.3B). This is because the simulated activity during Y_1 and Y_2 were more similar to each other than they were to simulated activity during X (by design).

We also applied this analysis to simulated data where firing rates represented only time, and did not explicitly represent the contingency. To generate firing rate simulations which represented only time and not contingency, we sorted the firing rates of each cell across a given session, such that a random half of the cells steadily increased their firing rates over the course of the session, and the other half decreased their firing rates over the course of the session. This ensured that we generated simulations which had the same number of cells and sessions as our actual data, as well as identical firing rate distributions as our actual data, but had firing rates which represented only the passage of time. Projections of simulated firing rates representing only time information did not overlap at all, and projections during Y_1 and Y_2 were more dissimilar from each other than from projections during X (Figure 4.3C).

What would these projections look like if the inputs were purely noise? To generate shuffled firing rates (Figure 4.3D,G), we shuffled the inter-spike intervals for each

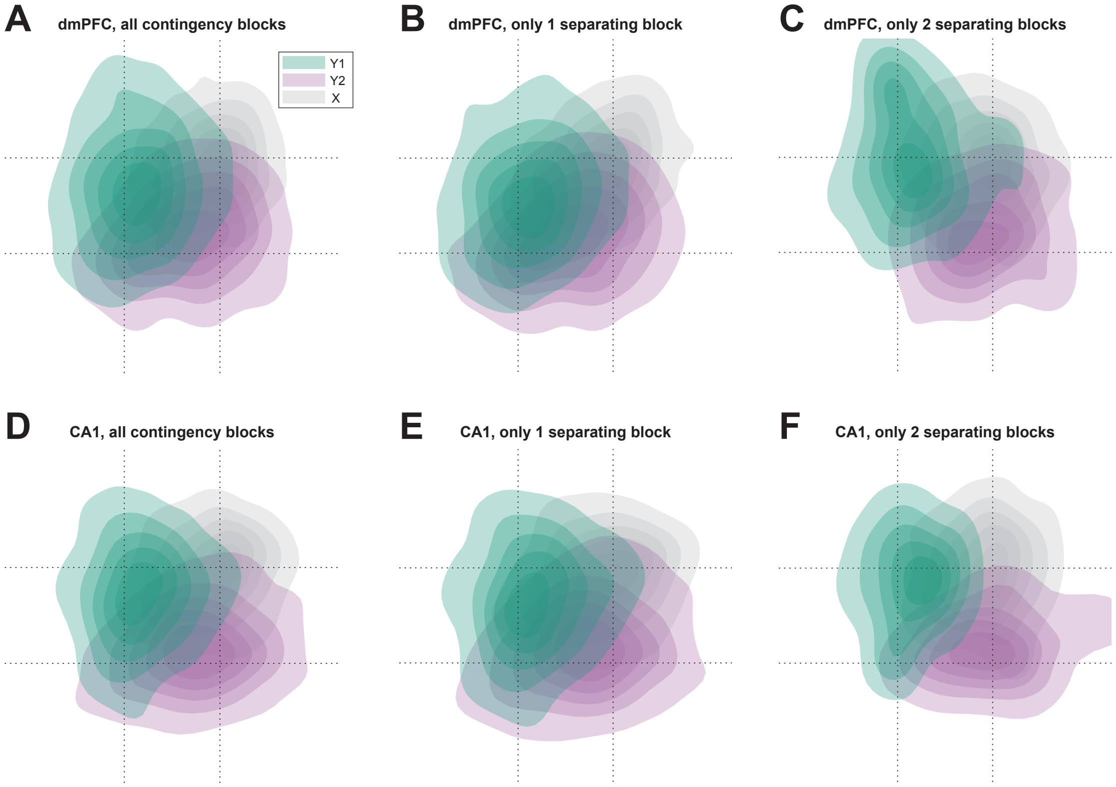

Figure 4.5: LDA projections for (A-C) dmPFC and (D-F) CA1 (as in Figure 4.3) for (A and D) all blocks, (B and E) for only $Y_1/X/Y_2$ contingency block sequences where X contained a single contingency block, and (E and F) for only $Y_1/X/Y_2$ contingency block sequences where X contained two contingency blocks.

cell independently, and re-generated that cell's spike times from the shuffled ISIs. This maintained some firing rate statistics (i.e. the mean firing rate of each cell) while removing any relationship to either contingency or time. The projections of randomly shuffled neural activity had projections which extensively overlapped for all three epochs (Figure 4.3D).

Projecting ensemble firing rates in dmPFC and CA1 onto LDA axes in this way revealed that neural activity in both structures showed signs of both contingency representation and a drift over time (Figure 4.3E,F). The separation between projections during either Y_1 or Y_2 and X was significantly larger than would be expected by chance (Figure 4.3G; 1000/1000 shuffles had lower Cohen's d between X and both Y projections, on both LDA axes in both structures). This suggests that neural encoding changed between subsequent contingency blocks.

To determine whether this difference was due to an encoding of contingency identity, or simply due to an unrelated change in ensemble encoding over time, we compared the projections of neural activity to the simulations encoding only contingency information or only time information. LDA projections of neural activity during Y_1 and Y_2 were more similar to each other than the projections of simulations representing only time, and were more similar to each other than to the projections of neural activity during X epochs (Figure 4.3H, 1000/1000 simulations had median projections further from the other Y block than was observed in the neural data for both LDA axes in both dmPFC and CA1). This suggests that the separation was in part due to representation of the contingency identity. However, not all of the representational change could be explained by contingency encoding: the LDA projections of neural activity during Y_1 and Y_2 were more separated than were the projections of simulations representing only contingency information (Figure 4.3H, 1000/1000 simulations had median projections closer to the other Y block than was observed in the neural data for both LDA axes in both dmPFC and CA1). This indicates that while contingency identities were represented, there was also a change in the ensemble activity over time which could not be explained by contingency representation.

To determine if any single specific transition type was primarily responsible for these results, we repeated the LDA projection analysis for each of the six switch types individually, but found that the vast majority of projections for each individual switch type were similar to that of the pooled data (Figure 4.4). Also, to account for whether the number of intervening contingency blocks was driving this effect, we repeated the LDA projection analysis for contingency block triplets containing only 1 intervening block, and those containing only 2 intervening blocks. The projections for each were again consistent with the projections when using all the data, though the projections for triplets with two intervening blocks were more separated, further suggesting the presence of a drift over time (Figure 4.5). These results demonstrate that ensemble activity in dmPFC and CA1 represented the abstract task rule or behavioral strategy, while simultaneously changing their representations over time in ways unrelated to contingency.

4.4 CA1 and dmPFC ensembles encoded the current contingency

4.4.1 Bayesian decoding of contingency

To investigate how strongly dmPFC and HPC encoded the task rule, and when these representations changed, we used Bayesian decoding to decode the current task contingency (Left, Right, or Alternate) from ensemble firing rates in CA1 or dmPFC.

We used Bayesian decoding (Zhang et al., 1998) to decode both spatial position and contingency from firing rates in either dmPFC or CA1 (Figure 4.6). We used decoding time bins of 100ms, and a 16×16 grid for spatial location, and 3 bins for the 3 contingencies. We used 100-fold cross validation to perform the decoding. To compute the posterior probability of a given contingency on a given lap (as opposed to only during a single time bin), we computed the cumulative log posterior probability for each contingency across time samples during which the rat was on the central segment of the maze during that lap.

Unlike the previous analysis, this Bayesian decoding analysis captured both spatial information and contingency information, and did not simply depend on the

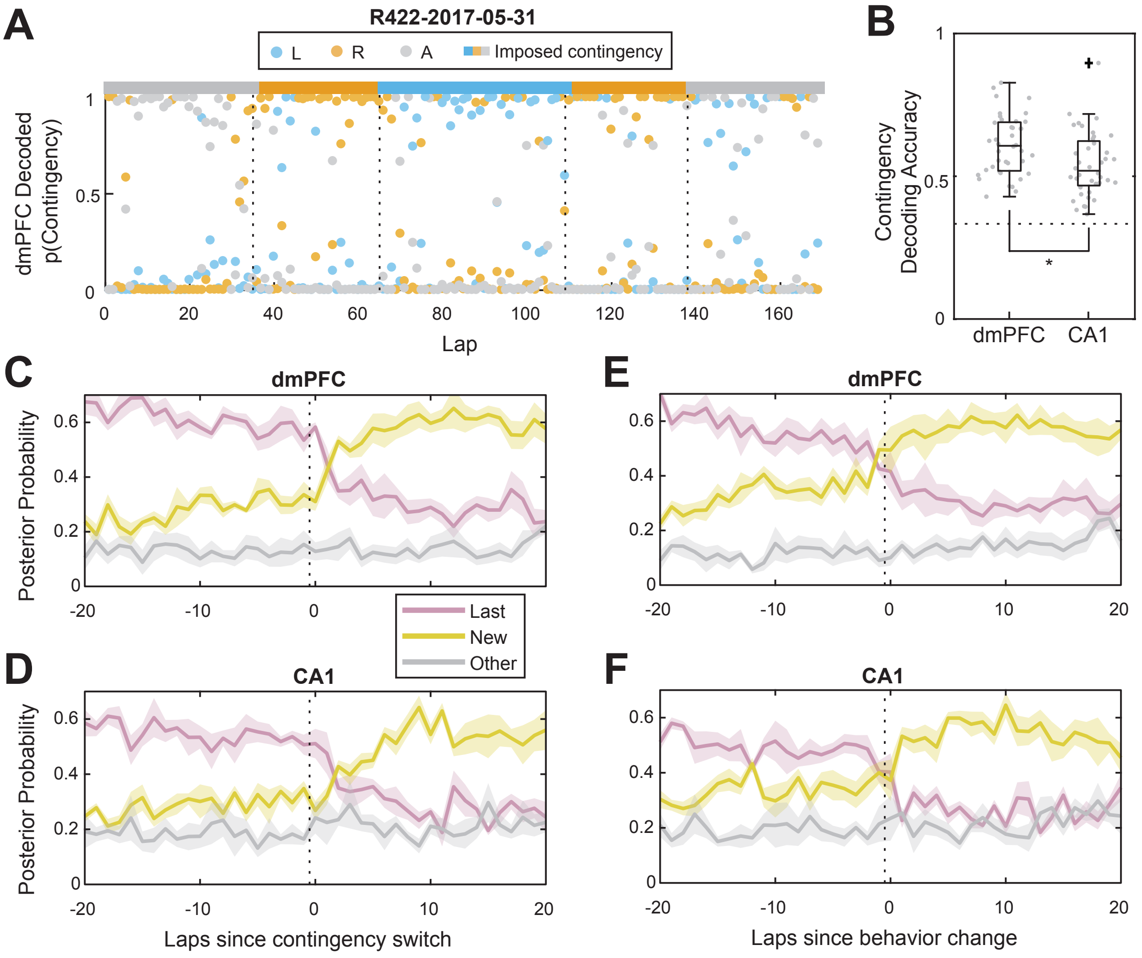

Figure 4.6: Bayesian decoding of contingency representations in dmPFC and CA1. (A) Decoding from dmPFC ensemble activity over an example session. Dots are per-lap decoding posterior probabilities, and the colored bar at the top indicates the imposed contingency. (B) Decoding accuracy from ensembles in dmPFC and CA1. $N = 40$ sessions. Dotted line corresponds to chance. (C) Decoding aligned to contingency switches for dmPFC, and (D) CA1. Dotted lines indicate the last lap of the previous contingency block. (E) Decoding aligned to behavioral change laps for dmPFC, and (F) CA1. Dotted lines indicate the last lap of the previous behavioral strategy. (C-F) show mean ± SEM, $N = 6$ rats.

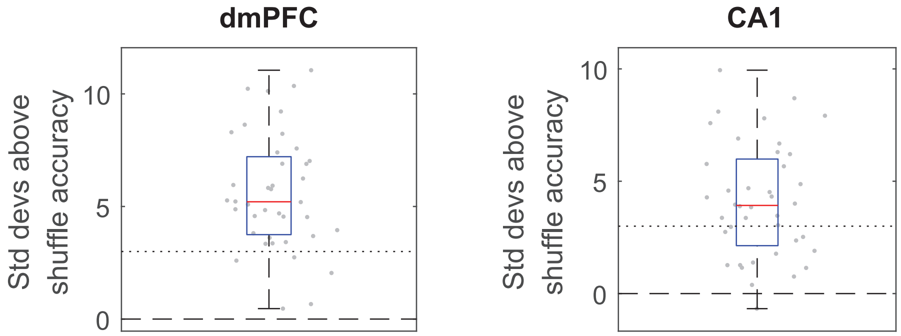

Figure 4.7: Contingency decoding from ensemble activity in dmPFC and CA1. Standard deviations above accuracy distributions resulting from shuffled spiketimes.

average firing rate of cells across a trial. Both dmPFC and CA1 encoded the current contingency more strongly than chance (Figure 4.6B, two-sided Wilcoxon signed rank test vs 1/3, $p = 3.6 \times 10^{-8}$ for both dmPFC and CA1, $N = 40$ sessions). 35 out of 40 sessions for dmPFC and 26 out of 40 sessions for CA1 had decoding accuracy greater than 3 standard deviations above the accuracy of decoding performed on shuffled firing rates (Figure 4.7, 100 shuffles per session).

Contingency decoding from ensembles in dmPFC was significantly more accurate than decoding from ensembles in CA1 (Figure 4.6B, two-sided Wilcoxon rank sum test $p = 0.0045$, $N = 40$ sessions).

Within around five laps after a contingency switch, ensembles in both dmPFC and CA1 began to represent the new contingency more strongly than the contingency from the previous block (Figure 4.6C,D). On average, ensembles in dmPFC represented the new contingency more strongly than the previous contingency before the animal updated its behavior to be consistent with the new contingency (Figure 4.6E), while in contrast this transition in CA1 was not different from the behavioral change point (Figure 4.6F and 4.11B).

4.4.2 Ensembles did not overtly remap

Did ensembles remap between contingency types? Other work finds that when environments change drastically, place cells sometimes completely remap – that is, they completely change their firing fields as if the animal was placed in an entirely new environment.

To compare the amount of global remapping occurring between contingency types, we measured the correlation of spatial tuning curves for the entire ensemble between contingency blocks. For each contingency block, we computed the average firing rate of each cell in each of 5 spatial bins along the central segment of the maze (from lap start, at the rear of the maze, to choice point entry). We then took these per-contingency-block firing rate vectors and computed the Pearson's correlation coefficient between blocks (Figure 4.8). Global remapping would have led to much lower correlation values between contingency blocks of a different type than between contingency blocks of the same type.

However, neither ensembles in CA1 nor dmPFC appeared to be overtly remapping between contingency types (Figure 4.8), suggesting the encoding of contingency may have been via more subtle changes in firing rate, such as rate modulation.

4.4.3 Ensemble activity was more correlated within-contingency

To measure the similarity between population activity before and after contingency switches, we correlated neural tuning curves between pairs of laps around the switches (Figure 4.9A,B). For each cell, we computed the firing rate in each of 5 spatial bins along the central segment of the maze (from lap start to just before choice point entry). For each pair of laps, we computed the Pearson correlation between the firing rate

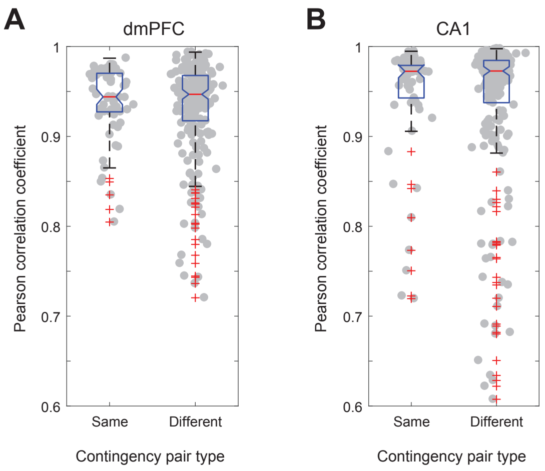

Figure 4.8: Ensembles do not appear to overtly undergo global remapping between contingency types. Pearson correlation between spatial tuning curves during passes through the maze's central track of ensembles in (A) dmPFC and (B) CA1.

vectors for the lap (each of which had $5 \times N_c$ elements, where N_c is the number of cells recorded on that day). We averaged windows of 20 laps on either side of contingency switches to generate a switch-aligned average correlation matrices for both dmPFC (Figure 4.9A) and CA1 (Figure 4.9B). We excluded laps before the previous switch or after the subsequent switch from this analysis, such that the data in Figure 4.9 only reflected correlations between laps in identical or adjacent contingency blocks.

Ensemble firing rates in dmPFC and HPC were more correlated within contingency blocks than between contingency blocks, and the appearance of this within-contingency correlation appeared to occur more slowly after a contingency switch in CA1 than in dmPFC (Figure 4.9).

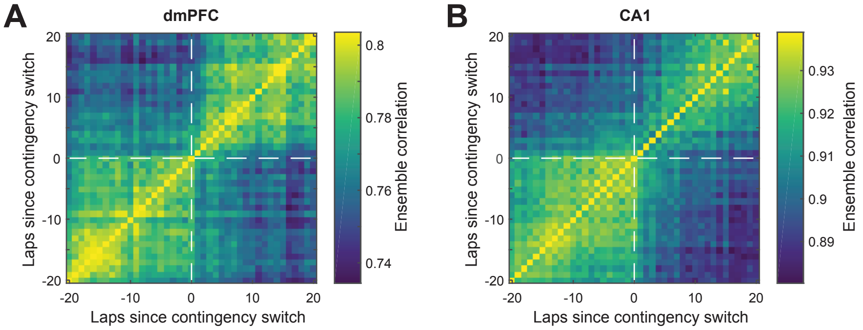

Figure 4.9: Ensemble correlations in dmPFC and CA1 aligned to the contingency switch. (A) Switch-aligned correlation matrix of firing rates in dmPFC. (B) Switch-aligned correlation matrix of firing rates in CA1.

4.5 Ensembles in dmPFC transitioned before CA1 ensembles

To quantitatively measure the timing of these representational transitions on a switch-by-switch basis, we performed a change-point analysis on the ensemble firing rates in dmPFC and CA1. We performed a clustering-based transition point analysis to determine on what lap neural representations were most likely to transition from one representation to another (Powell and Redish, 2016). We used K-means clustering to cluster neural activity on each lap (again split into 5 spatial bins, only during the central segment of the maze, as in the correlation analysis above) into 3 clusters (for the 3 possible contingencies). We then applied the change point analysis from Gallistel et al. (2004) to the cluster IDs of each lap. Because K-means assigns clusters stochastically, we repeated this analysis 1000 times for each session to obtain a probability distribution across laps within the session, which captured the probability of neural activity transitioning from one state to another.

The firing rate vector for each lap contained not just each cell's average firing rate across the lap, but the average firing rate of each cell in each of 5 spatial bins across

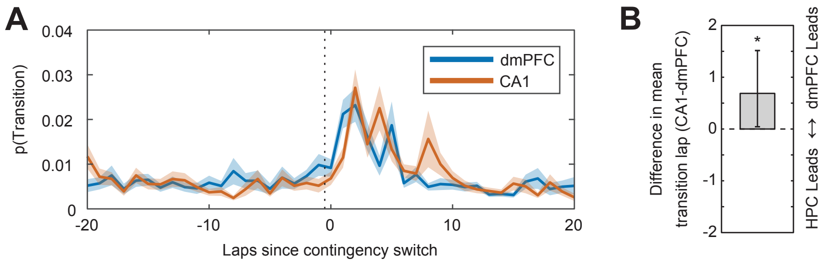

Figure 4.10: Representational transitions after contingency switches in both dmPFC and CA1. (A) Transition probability for dmPFC and CA1 aligned to contingency switches. Lines and shaded areas show mean ± SEM, $N = 164$ switches. (B) Median difference of the transition probability distributions and bootstrapped 95% confidence interval. $N = 164$ switches.

the central maze path from the beginning of a lap to the choice point. Because there is a level of stochasticity inherent in k-means clustering (a different initialization can result in a different clustering), we repeated the clustering and transition detection procedure many times to obtain a probability distribution of ensemble transitions over laps (Powell and Redish, 2016).

To compute the timing difference between dmPFC and CA1 (Figure 4.10), we took the difference between the means of the transition probability distribution in dmPFC and CA1 for each switch, and computed bootstrapped 95% confidence intervals on the median.

This analysis revealed that representations in both dmPFC and CA1 ensembles very likely underwent transitions within a few laps after contingency switches (Figure 4.10A). The transition in dmPFC occurred significantly ahead of the transition in CA1 (paired two-sided Wilcoxon signed rank test comparing the means of the per-switch transition probability distributions, $p = 0.039$, $N = 160$ contingency switches). The median lead by dmPFC was 0.69 laps (Figure 4.10B, 95% confidence interval between 0.045 and 1.5 laps). However note that our analysis was unable to resolve changes on sub-lap timescales, so we were only able to determine that the transition in dmPFC

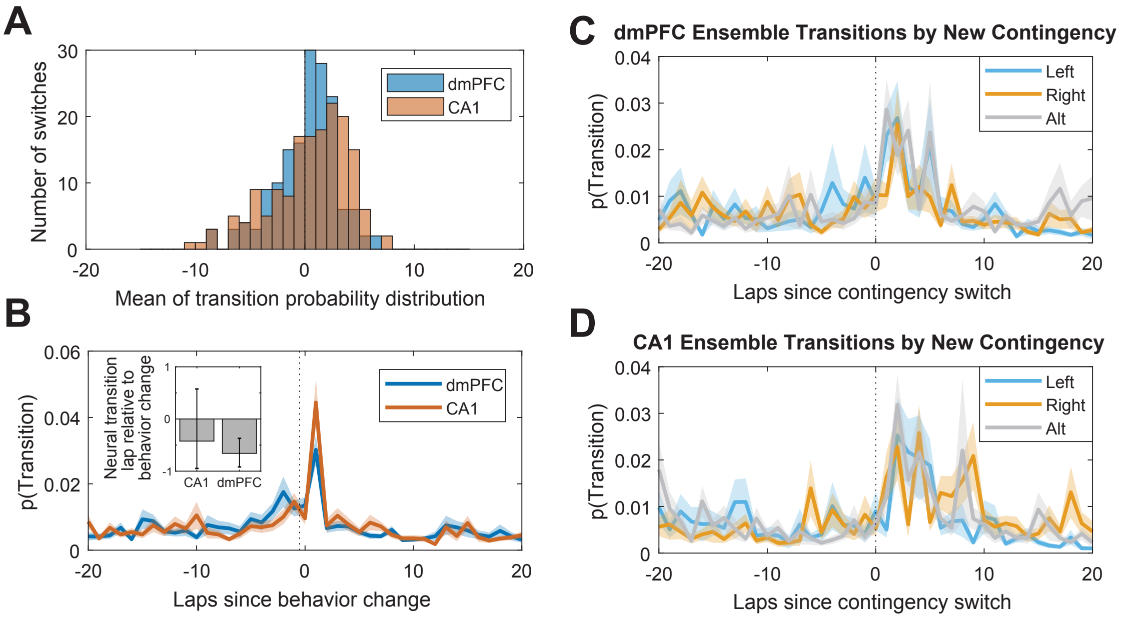

Figure 4.11: The timing of representational transitions in dmPFC and CA1 relative to behavior. (A) Means of ensemble transition probability distributions relative to contingency switches for dmPFC and CA1. (B) Transition probability aligned to the behavior change. Shown is mean ± SEM, $N = 164$ contingency switches. Inset shows the mean neural transition lap relative to the behavioral change, with bootstrapped 95% confidence intervals. $N = 164$ contingency switches. (C) Transition probability split by new contingency in dmPFC. (D) Transition probability split by new contingency in CA1.

preceded that in CA1, but not by exactly how much. Importantly, the repeated clustering is not what provided statistical power for this analysis, but was used only to ensure that the resulting mean cluster transition lap was less biased by the k-means initialization. Rather, the timing difference between cluster transitions was evaluated using Wilcoxon rank sum tests.

In both dmPFC and CA1, the neural ensemble transition was most likely to occur on the lap when rats updated their behavioral strategies to be consistent with the new contingency, but dmPFC ensembles were significantly more likely to transition before the behavioral change, while the transition in CA1 did not occur at a significantly different time from the behavioral transition (Figure 4.11B). There did not appear to be any major differences in the transition time course depending on the identity of the new contingency in either dmPFC or CA1 (Figures 4.11C,D). However, in both dmPFC and CA1 there was a correlation between the amount of VTE and the probability of a neural transition (Two-sided Wilcoxon sign rank test $p = 0.020$ in dmPFC and $p = 0.0066$ in HPC, $N = 40$ sessions, of per-session Spearman's rank correlation coefficients between zIdPhi and neural transition probabilities), suggesting that these brain areas were more likely to update their contingency representations on laps where rats deliberated.

4.6 Representations in dmPFC drifted faster than in CA1

Comparing the projections of neural activity to simulations representing only time (Figure 4.3) suggested that representations changed over time, but could not tell us how quickly. To investigate the speed at which neural representations in dmPFC and CA1 changed across time, without including changes due to the type of contingency, we measured the Pearson correlation of ensemble spatial tuning curves between pairs of contingency blocks of the same type, as a function of the time separating the blocks. To compute the speed of correlation decay, we divided the correlation coefficient by the number of laps which separated the centers of the two contingency blocks.

Ensemble activity was less correlated between pairs of contingency blocks which were further apart in both dmPFC (Figure 4.12A) and CA1 (Figure 4.12B). The ensemble correlation coefficient decreased by 6.7×10^{-4} per lap in dmPFC (Figure 4.12C, 95% confidence interval on the median was between -8.5×10^{-4} and $-5.3 \times$

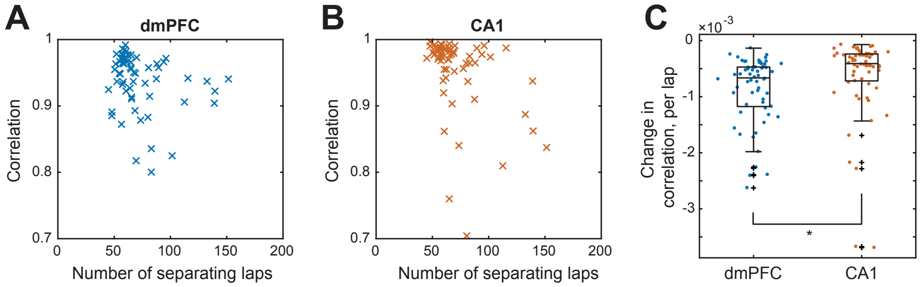

Figure 4.12: The speed of the representational drift over time. (A) Correlation between ensemble tuning curves as a function of how many laps separate the blocks being correlated, in dmPFC and (B) in CA1. (C) The change in the correlation per lap for both dmPFC and CA1. $N = 62$ pairs of contingency blocks of matching types.

10^{-5}), while in CA1 the ensemble correlation coefficient decreased by 4.1×10^{-4} per lap (Figure 4.12C, 95% confidence interval on the median was between -4.7×10^{-4} and -2.9×10^{-4}). The decrease per lap in the correlation coefficient was greater in dmPFC than in CA1 (Figure 4.12C, Two-sided Wilcoxon rank sum test $p = 6.88 \times 10^{-5}$, $N = 62$ contingency pairs). This suggests that the change in the representation of non-contingency time-varying information, or representational drift, occurred more quickly in dmPFC than in CA1.

To validate that this drift was not due simply to the physical drift of our recording electrodes over time, we compared the spike waveform self-similarity of identified units across time to the similarity between different units, similar to the analysis used by (Tolias et al., 2007) to validate stable recordings across days. We computed the Euclidean distance of spike waveforms between the first and second halves of the session during which that unit was recorded. We also computed the distances between the average waveforms of each unit and waveforms of other, non-identical, units within the same session. The distances between identified units across time was far lower than the distances between each unit and different units (Figure 4.13). This indicates that our recordings suffered minimally from electrode drift, and thus

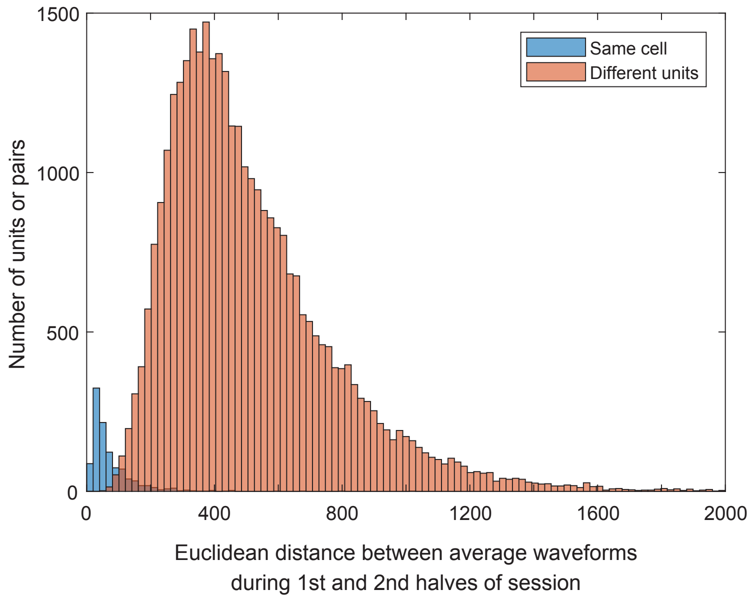

Figure 4.13: Distance of average spike waveforms from identified single units (blue) between the first half and second half of the session during which that unit was recorded. Also, the average spike waveform distance of unit pairs which were identified as different units during the same session (orange). Identified units had waveforms which were much more self-similar than to control unit waveforms, suggesting minimal recording drift. This analysis is similar to, but not identical to, the analysis performed in Tolias et al. (2007) to identify cell stability across days.

electrode drift was unlikely to explain our observation of firing rate changes over time, suggesting that actual representational change accounted for the observed changes in firing rate patterns across time.

4.7 Discussion

Dorsomedial prefrontal cortex is thought to exert top-down contextual control on hippocampal spatial encoding, but the timecourse and dynamics of this interaction are unknown. We recorded from ensembles in dmPFC and CA1 simultaneously on a task with multiple rule switches, and found that both dmPFC and CA1 represented task contingencies while concurrently representing other information which changed over time. Representations in dmPFC changed faster than in CA1, and dmPFC began to represent new task rules before HPC. Our results suggest that top-down information from dmPFC about contextual information, such as information about task rules, appears first in dmPFC and likely alters representations in HPC. The fact that the representational transition we observed in dmPFC preceded the transition in HPC suggests that this information may not immediately be incorporated into hippocampal representations, perhaps due to an inherent stability of hippocampal representations.

Given the fact that firing rates do change over time, when inspecting the neural encoding of information yoked to blocks of time, it becomes especially important to perform some sort of representational stability analysis (Figure 4.3) to confirm that the decoding of this information is not simply an artifact of representational drift. The analysis we performed here is not the only option: any probabilistic classification model can be trained to separate neural activity during one presentation of a given contingency from activity during between epochs, and then tested to see how well it classifies activity during a subsequent presentation of the contingency of interest. We opted for linear discriminant analysis both for its simplicity and because it lends itself to the more visually interpretable approach we took here (two dimensional projections, as opposed to comparing classification metrics).

Although representations are known to drift over time (Mankin et al., 2012; Hyman et al., 2012; Ziv et al., 2013), it is unknown what causes this drift, or what purpose it serves. One explanation is that the drift we observed could have been simply due to the representation of additional information which was changing over time, such as signals related to satiety or motivation. However, theories of how the brain encodes temporal context (Mensink and Raaijmakers, 1988; Howard and Kahana, 2002) suggest that drifting representations could facilitate the retrieval of memories which are closer in time, via associative dynamics. The temporal context model of Howard and Kahana (2002) proposes that the drift is not random, but rather is driven by the retrieval of recent contextual information.

In a related study, Malagon-Vina et al. (2018) observed on a rule-switching task that strategy or rule representations in dmPFC differed between the first and later repeated rule presentations. They concluded that a new rule representation occurs each time a previously-presented rule occurs, suggesting that dmPFC may be encoding only rule changes, and not the actual rule identity. While it may appear that their conclusions are in conflict with our results that there are stable rule representations in dmPFC, we believe that our conclusions are actually consistent with those of Malagon-Vina et al. (2018). We observed that while there is a stable representation of rule or strategy identity, there is simultaneously a drift in the representation over time, due to any number of other factors (such as satiation or motivation). We hypothesize that this representational drift over time - and not any inherent change in contingency encoding - causes the apparent representation of rule identity to differ between the first and repeated presentations of the rule, which may have led Malagon-Vina et al. (2018) to conclude that a new representation is generated each time a rule is repeatedly presented. Of course this is somewhat of a semantic issue: the ensemble firing rates did indeed change between rule presentations. However, we find that even with the change in representation over time, there remains within dmPFC a consistent underlying representation of contingency.

Our data shows that both dorsomedial prefrontal cortex and hippocampus encode contextual information about the current contingency, while simultaneously encoding other information which changes over time throughout the task. Furthermore, our results suggest that these context representations are more static in some brain areas, such as hippocampus, while they are more dynamic in others, such as dorsomedial prefrontal cortex.

Chapter 5

Interactions between dmPFC and CA1 during Deliberation

5.1 Introduction

In the previous chapter we examined contingency representations in dorsomedial prefrontal cortex and hippocampus individually, and how those representations changed between laps. We found that hippocampal representations of contingency took longer to update than in dmPFC, a difference which was on the order of laps, not milliseconds. However, we know that information transfer between these two areas could in theory occur much faster. Work examining the relationships between local field potentials (LFPs) in these two areas finds coherence between the LFPs in PFC and HPC, specifically in the theta frequency band (6-11Hz), which suggests interactions on subsecond timescales (Siapas et al., 2005; Jones and Wilson, 2005b; Hyman et al., 2010; Colgin, 2011; Gordon, 2011; O'Neill et al., 2013; Brincat and Miller, 2015). Furthermore, there are both monosynaptic and polysynaptic projections from CA1 to dmPFC (Swanson, 1981; Ferino et al., 1987; Jay and Witter, 1991; Verwer et al., 1997; Delatour and Witter, 2002; Floresco and Grace, 2003; Hoover and Vertes, 2007), as well as a bisynaptic and bidirectional connection between CA1 and dmPFC via the nucleus reuniens and other thalamic nuclei (Vertes, 2002, 2004; McKenna and Vertes, 2004; Vertes et al., 2006; Di Prisco and Vertes, 2006; Vertes et al., 2007; Hoover and Vertes, 2012; Cassel et al., 2013; Dolleman-Van der Weel et al., 2017; Dolleman-van der Weel et al., 2019), so it is entirely possible that these two areas are able to transfer information within timescales on the order of tens of milliseconds.

However, as we saw in the previous section, contingency information does not appear to transfer from dmPFC to HPC on timescales this fast. What then, if any, information is shared between the two structures on faster timescales? Theories of the deliberative system suggest that prefrontal areas initiate simulation of potential actions, and then keep track of the estimated outcomes of those internal simulations. These events are thought to occur on fast timescales, and so reasonable candidates for information shared between dmPFC and CA1 on faster timescales include information about candidate actions (choice), spatial information, and information about reward or value. In this chapter we examine interactions between HPC and dmPFC on these faster timescales, and specifically investigate how spatial information and representations of reward are related between the two structures.

5.2 Coherence between dmPFC and CA1 Local Field Potentials

If dorsomedial prefrontal cortex and hippocampus interact during decision-making, we would expect this interaction to be reflected in a relationship between the local field potentials in each structure. Previous work has shown that during working memory tasks the local field potentials in PFC and HPC become synchronized, specifically in the theta frequency band, which is usually around 6-11Hz in rats (Siapas et al., 2005; Jones and Wilson, 2005b; Hyman et al., 2010; Colgin, 2011; Gordon, 2011; O'Neill et al., 2013; Brincat and Miller, 2015). Furthermore, the timing between oscillations in dmPFC or HPC depends on whether the animal is recalling information at decision time or encoding information during exploration (Place et al., 2016; Liu et al., 2018).

Some work also finds relationships in other frequency bands including gamma (30-80Hz) and delta (∼2-4Hz). Gamma oscillations in the prefrontal cortex are phase-modulated by the hippocampal theta rhythm (Sirota et al., 2008). Local field potentials in the two structures have even been found to synchronize at gamma frequencies (Spellman et al., 2015), and also at the far slower delta band (Fujisawa and Buzsáki, 2011). There appears to be especially strong theta coherence between prefrontal and the hippocampus during moments of tasks requiring working memory (Colgin, 2011; Gordon, 2011; Fujisawa and Buzsáki, 2011) such as at the choice points of spatial tasks during decision making (Benchenane et al., 2010).

Here, we investigated coherence between theta oscillations in dorsomedial prefrontal cortex and local field potentials recorded from the hippocampal fissure. To analyze power and coherence we used the Chronux software package for Matlab (Mitra et al., 2018; Mitra, 2007). We found that local field potentials in both the hippocampus and dmPFC displayed strong theta oscillations, peaked at around 8Hz (see Figure 5.1A for an example). We also observed a strong beta component of the LFPs (12-25Hz), although this is difficult to parse out from simply a theta harmonic. It has long been known that theta power in the hippocampus increases with running velocity in rats. We found increased theta power during the run segments of the maze in not only HPC but also in dmPFC (Figure 5.1C). We found strong coherence between dmPFC and HPC in the theta band, and also in the beta band (Figure 5.1B,D). Again, we

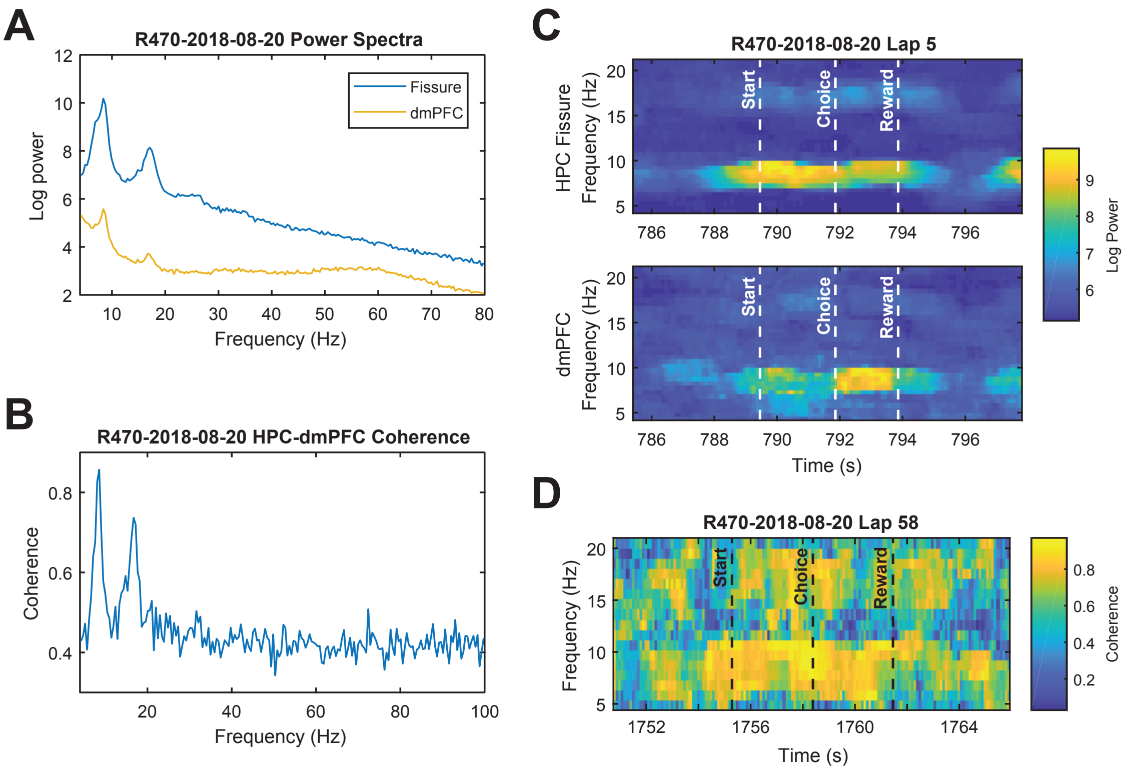

Figure 5.1: Examples of power and coherence between dmPFC and hippocampal LFPs. (A) Example power spectrum in the hippocampal fissure and in dmPFC from one session. (B) Example coherence between HPC and dmPFC for the same session. (C) Example spectrogram across one lap for both HPC and dmPFC. Vertical dotted lines indicate the start of the lap, choice point entry ("Choice"), and reward zone entry ("Reward"). (D) Example coherogram across the same lap.

suspect this "beta-band" peak in power and coherence in our data is simply a theta harmonic or due to the non-sinusoidal shape of theta oscillations.

Theta coherence between dmPFC and hippocampus was greater while rats ran the central maze segment than before they began that lap ($p = 1.8 \times 10^{-5}$, Wilcoxon rank sum test, $N = 40$ sessions, Figure 5.2A-C, and also see Figure 5.1D for an example). However, theta coherence was greater still as rats passed through the choice point of the maze ($p = 9.5 \times 10^{-5}$, Wilcoxon rank sum test, $N = 40$ sessions, Figure 5.2A-C). These findings are consistent with previous work which finds higher theta coherence between PFC and HPC at choice points on working memory-dependent tasks (Benchenane et al., 2010).

Is coherence between HPC and dmPFC related to vicarious trial and error? If the theory is correct that theta coherence between HPC and PFC arises because the two structures are working together to retrieve information being held in working memory, then we might expect to see an increase in coherence at choice points while rats display VTE. On the other hand, rats pause during vicarious trial and error events (Muenzinger and Gentry, 1931; Redish, 2016), and theta power is known to be related to running speed, and therefore theta coherence might be lower during VTE events simply because theta power is lower than it would be had the rat not paused.

We found that theta coherence between dmPFC and HPC decreases during passes through the choice point where rats displayed vicarious trial and error (Figure 5.2D). We also observed that rats' velocity was lower during passes through the choice point on laps which they made errors (Figure 5.2F). This is probably because rats displayed post-error slowing on the contingency switch task (section 3.2), and were more likely to display vicarious trial and error on laps after contingency switches (Figure 3.4C), when they were also far more likely to be making errors (Figure 3.4A) due to not yet having figured out the new contingency. While theta coherence between dmPFC and HPC was lower at the choice point during passes where rats displayed VTE, it was still higher than at other moments where rats paused, such as while waiting for food reward (Figure 5.2D, compare pre- or post-lap coherence to that at the choice point during VTE).

Some work reports there is greater synchrony between PFC and HPC in the theta band following errors (Brincat and Miller, 2015). We found that indeed theta coherence between dmPFC and HPC was significantly higher during the 5s after reward zone entry following incorrect choices than following correct choices ($p =$

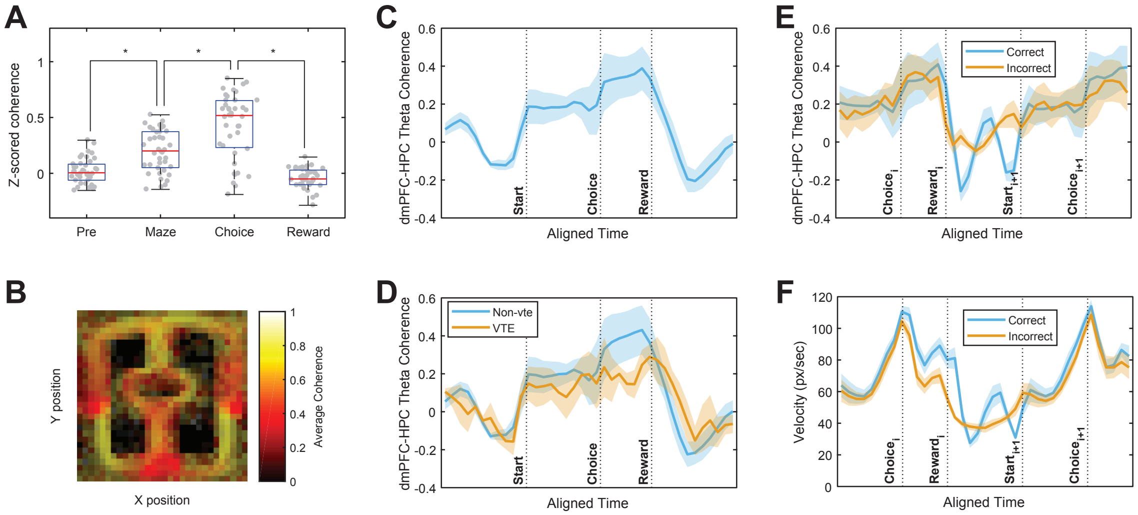

Figure 5.2: Theta coherence between dmPFC and HPC. (A) z-scored theta coherence during a 5s period before lap start ("Pre"), during running of the central maze segment ("Maze"), during passes through the choice point ("Choice"), and during a 5s period after reward zone entry ("Reward"). $N = 40$ sessions. (B) Average coherence as a function of location on the contingency switch task maze. (C) z-scored theta coherence as a function of linearized position through a lap. Vertical dotted lines show lap start ("Start"), choice point entry ("Choice"), and reward zone entry ("Reward"). (D) same as in (C), split by whether VTE occurred on that lap. (E) z-scored theta coherence over the course of two sequential laps, split by whether the rat's choice on the first lap was rewarded or unrewarded. Vertical lines indicate choice point entry on the first lap ("Choice$_i$"), reward zone entry at the end of the first lap ("Reward$_i$"), start of the second lap ("Start$_{i+1}$"), and choice point entry on the second lap ("Choice$_{i+1}$"). (F) Rat velocity over the course of two sequential laps, split by whether the rat's choice on the first lap was rewarded or unrewarded. Shaded areas in C-F show mean ± standard error, $N = 6$ rats.

1.3×10^{-5}, Wilcoxon signed rank test, $N = 40$ sessions). However, rather than an increase in coherence following errors, a more accurate description is perhaps that coherence decreases less following errors (Figure 5.2E). That is, upon reward zone entry, theta coherence between HPC and dmPFC still decreased following errors, but it decreased far more following correct choices. Also, this difference was not able to be fully explained by differences in running speed: the running velocity of rats following reward zone entry after errors was only slightly different than following correct choices (Figure 5.2F), while the coherence was drastically different (Figure 5.2E). Our results are consistent with previous work finding that coherence is greater after errors (Brincat and Miller, 2015), but highlights that on this task this difference is due to a lesser decrease in coherence following errors, and not because of a spike in coherence.

5.3 Correlation between Non-local Spatial Representations

In the previous section we found that theta oscillations in HPC and dmPFC were coherent, but how was the information represented in these two areas related? We performed cross-validated Bayesian decoding of location from ensemble spiking activity in dmPFC, and also separately for simultaneously recorded ensembles in HPC. Using the decoded spatial posterior distributions, we computed how far ahead of or behind rats' actual positions HPC and dmPFC represented, and how the spatial representations in each area were related (Figure 5.3A).

Previous work has found that HPC represents locations further from the rat's actual location during vicarious trial and error (Johnson and Redish, 2007). Replicating this previous work, we found that during passes through the choice point on the contingency switch task, indeed HPC encoded positions further ahead of the rat during VTE than during non-VTE choice point passes ($p = 0.0156$, one-sided Wilcoxon signed rank test, $N = 6$ rats, Figure 5.3C). Similarly, the dmPFC also encoded positions further ahead of the rat during VTE at the choice point ($p = 0.0469$, one-sided Wilcoxon signed rank test, $N = 6$ rats), though the size of this effect was less prominent than that observed in HPC (Figure 5.3C). This suggests that both

HPC and dmPFC represented prospective spatial information during deliberation.

However, simply because both areas represented prospective information during deliberation does not necessarily mean they represented identical spatial information at the same time. To determine whether spatial representations in dmPFC and HPC were tightly locked on a fast timescale, we correlated the distance ahead or behind rats being represented by ensembles over the course of single theta cycles. We found that there was a small correlation between the relative position represented in HPC and the relative position represented in dmPFC (a mean Spearman's correlation coefficient of 0.0458, which was significantly greater than zero, $p = 0.0469$, one-sided Wilcoxon signed rank test, $N = 6$ rats, Figure 5.3B).

To determine whether these correlations were due to both areas representing goal locations simultaneously, as opposed to simply being a result of minute positional differences in the spatial representations across theta cycles, we analyzed the decoded spatial locations by what zone was being represented. Because the Bayesian decoding resulted in a decoding posterior across the entire maze, we were able to define three zones of interest: the choice point, the reward zone on the left side of the maze, and the reward zone on the right side of the maze (Figure 5.3D). To perform the following correlations, we computed the sum of the decoding posterior within each zone (normalized by area) per theta cycle. Encoding of the choice point versus either reward zone was correlated between dmPFC and HPC (a median Spearman's correlation coefficient of 0.129, which was significantly greater than zero, $p = 0.0313$, Wilcoxon signed rank test, $N = 6$ rats, Figure 5.3E). However the identity of the reward site being encoded by dmPFC and HPC was not significantly correlated ($p = 0.156$, Wilcoxon signed rank test, $N = 6$ rats, Figure 5.3F). The correlation coefficients for choice point and either reward zone representation were greater than the correlation coefficients for the identity of the encoded reward site (paired Wilcoxon signed rank

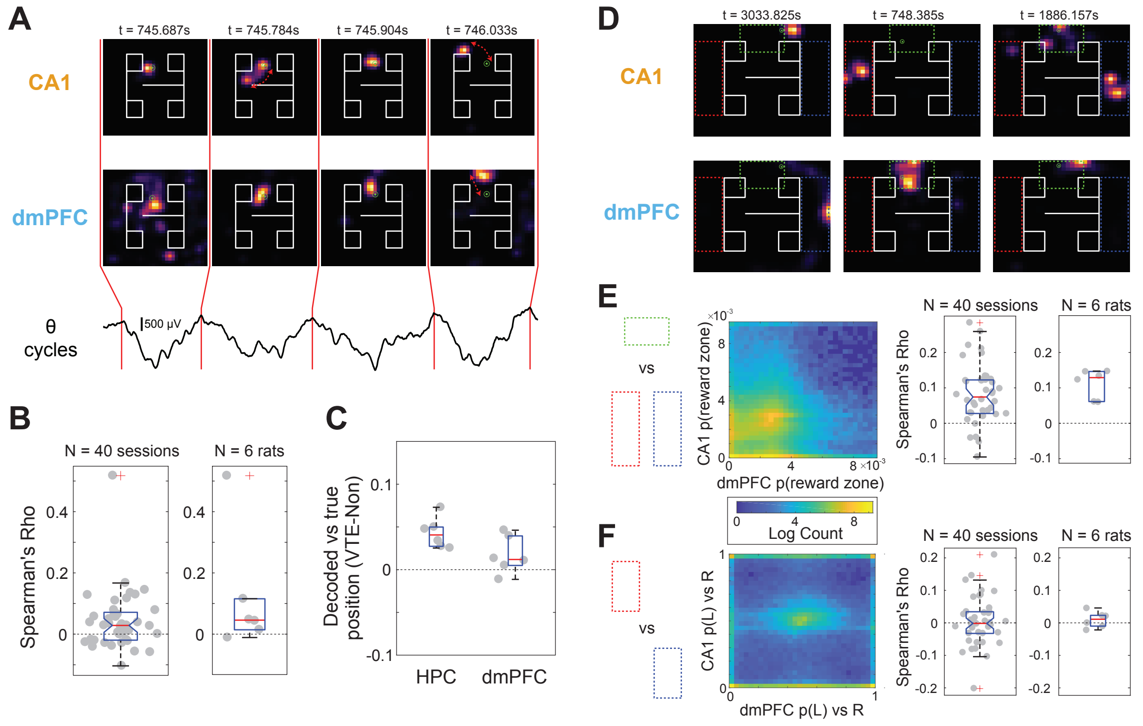

Figure 5.3: Correlations between location encoding in dmPFC and CA1. (A) Examples of Bayesian decoding of spatial location was performed on ensemble activity in dmPFC and CA1 using time bins corresponding to hippocampal theta cycles. Shown are example decoded posterior distributions from four consecutive theta cycles. Green dots indicate the actual position of the animal's head. Red arrows indicate discrepancies between the animal's true location and the decoded location. (B) Correlation coefficients of the distance relative to the rats true position decoded from ensemble activity between dmPFC and HPC, for each session and for each rat. (C) Difference in the average decoded position relative to the actual position of the rat at the choice point between VTE and non-VTE passes. Higher values indicate positions further ahead of the rat were represented during VTE passes. Units of the y-axis are laps (full revolutions around the maze). (D) Examples of Bayesian decoding and how it was used to compute the posterior probability of three different zones: the choice point (green dotted box), the reward zone on the left side of the maze (red dotted box), and the reward zone on the right side of the maze (blue dotted box). The top row shows decoding from CA1, the bottom row shows decoding from dmPFC, and each column shows decoding from the same theta cycle. (E) Correlation between dmPFC and HPC encoding of the choice point vs either reward site. (F) Correlation between dmPFC and HPC encoding of the reward site identity.

test, $p = 0.0313$, $N = 6$ rats, Figure 5.3E vs Figure 5.3F).

These results suggest that while both the dmPFC and HPC may have been in deliberative modes simultaneously, and both seemed to represent either local or prospective information within the same theta cycles, that they were not necessarily representing identical information at the same time. This is consistent with our results from section 4.5 concerning the timing of contingency representation transitions, which taken together with these results suggest that information from dmPFC may take some time to be incorporated into hippocampal representations, perhaps due to an inherent stability of hippocampal representations.

5.4 dmPFC Predicts Non-local Spatial Representation in HPC

If dmPFC instigates internal simulations and evaluations of candidate actions, carried out by the hippocampus and other structures, then neural signatures corresponding to that initiation should be present in dmPFC ensemble activity. Therefore, it should be possible to predict from ensemble activity in PFC whether hippocampus will represent prospective information.

To determine whether activity in dmPFC carried information about whether HPC was representing prospective information, we first used cross-fold Bayesian decoding to decode the location represented by HPC on a per-theta-cycle basis. We categorized hippocampal theta cycles while rats passed through the choice point as either "local" (the top 10% of theta cycles with the highest posterior density in the choice point) or "prospective" (the top 10% of theta cycles with the highest posterior density in the reward zones – either the left or right reward zone, as in Figure 5.3E). Then,

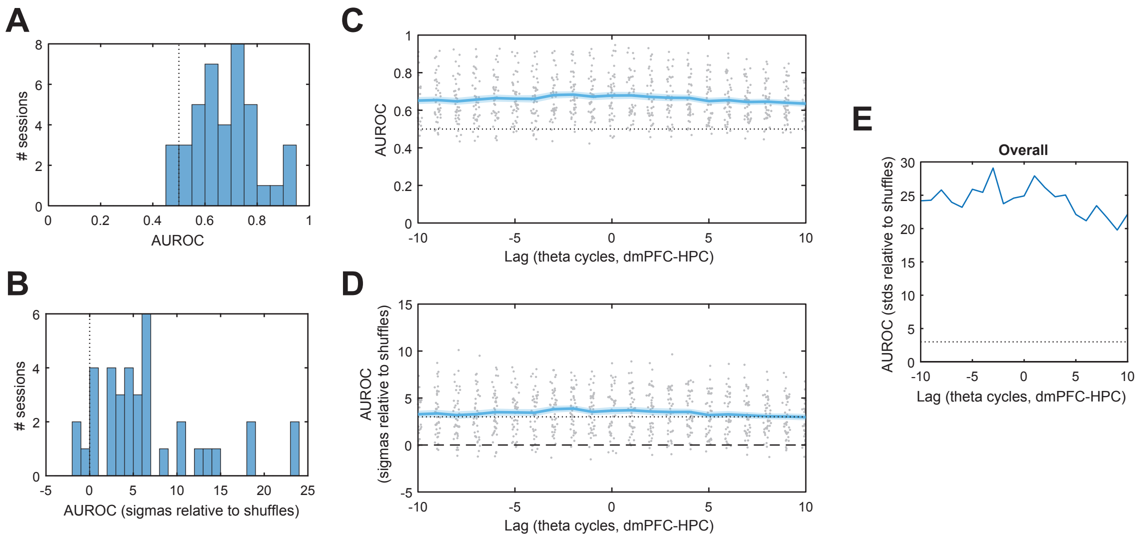

Figure 5.4: Predictions of prospective representation in HPC from activity in dmPFC. (A) Area under the receiver operating characteristic curve per session for the classifier predicting whether HPC is representing local or prospective spatial information, trained on dmPFC firing rates. (B) The AUROC relative to the AUROC of models fit to shuffled dmPFC firing rates. The x-axis is the number of standard deviations above the shuffle distribution. (C) AUROC of the classifier as a function of when dmPFC firing rates were used to predict representation in HPC. Negative values on the horizontal axis correspond to when dmPFC firing rates were used to predict upcoming representation in HPC, and positive values on the horizontal axis correspond to when dmPFC firing rates were used to predict past representation in HPC. Line and shaded area shows the mean ± standard error, $N = 40$ sessions, while the grey dots show the AUROC for each individual session. Horizontal dotted line shows 0.5, corresponding to the AUROC expected from noise. (D) Same as in panel C, but showing the standard deviations of the AUROC above the corresponding shuffle distributions as in panel B. The horizontal dotted lines show the median of the shuffle distribution and 3σ. (E) The AUROC relative to shuffle distributions for all sessions combined.

we trained a classifier (also a cross-fold Bayesian decoder) to predict from ensemble activity in dmPFC whether simultaneous hippocampal activity was representing local or prospective information. We measured the performance of the classifier using the Area under the receiver operating characteristic curve (AUROC) metric. An AUROC value of 0.5 indicates a classifier is performing at chance, while an AUROC value of 1.0 would indicate the decoder is correctly classifying every single theta cycle.

We found that this classifier trained on dmPFC activity performed well above chance, with AUROC values generally in the 0.6 to 0.8 range (Figure 5.4A). To ensure that the classifier was performing above chance, we repeatedly trained models on shuffled dmPFC firing rates (we used 1000 shuffles per session). The performance of the classifier trained on actual dmPFC firing rates was better than shuffles for all but 3 sessions, and was significantly above the shuffle distributions for the vast majority of sessions (Figure 5.4B).

However, the predictive power of dmPFC activity did not seem to be highly temporally specific. We repeated the classification analysis, but using neural activity from dmPFC at different time lags relative to the HPC theta cycle for which the classifier was trying to predict. If, say, activity in dmPFC were causing the next theta cycle in HPC to represent nonlocal information, but had no influence over hippocampal representations during preceding or subsequent theta cycles, then we would expect to see a sharp peak in the classifier's performance at a lag of -1 theta cycle. In contrast to that hypothesis, the performance of the classifier had a very broad performance profile as a function of lag (Figure 5.4C,D,E). This result suggests that if dmPFC influences hippocampal prospective representations, that the timecourse of this influence may be very diffuse – on the order of seconds or more, and not on a per-theta-cycle basis.

5.5 Reward Encoding in dmPFC

Our results in the previous section suggest that dmPFC could play some role in influencing hippocampal circuitry to enter into prospective modes, but it is also thought that hippocampal activity has an effect on prefrontal representations. The prefrontal cortex and the hippocampus are thought to form an information-processing loop, where top-down inputs from prefrontal cortex influence the retrieval of information from hippocampus, and that retrieved information informs prefrontal representations of contingencies, states, and potential actions (Wang et al., 2015; Jai and Frank, 2015; Redish, 2016; Shin and Jadhav, 2016; Eichenbaum, 2017). Generally, value-based decision making is thought to occur by comparing estimated values for each potential action, and taking the action with the highest expected value (Rangel et al., 2008; Padoa-Schioppa, 2011). More specifically, theories suggest that prefrontal structures may instigate internal simulations of outcomes of candidate actions by an internal model via brain areas including the hippocampus (Johnson and Redish, 2007; Hassabis and Maguire, 2009; Wang et al., 2015), that the value of the outcomes of those simulations are evaluated by other structures such as the ventral striatum (van der Meer and Redish, 2010; van der Meer et al., 2012), and that the valuations of the simulated outcomes of the candidate actions are used to select which action to perform.

If the prefrontal cortex uses estimates of reward or value associated with candidate actions to perform action selection, then presumably it or other areas must keep track of the candidate actions and the corresponding value estimates. However, there are different ways in which this might play out algorithmically, and different brain areas may play different roles in this process. One possibility is that the prefrontal cortex uses a similar mechanism for value-based decision making as sensory areas do for sensory-based decision making, by integrating evidence until some decision threshold is reached. These so-called “drift diffusion models” or “sequential sampling models” have been found to explain both behavior and neural activity in sensory areas during sensory decision-making tasks (Stone, 1960; Ratcliff, 1978; Hanes and Schall, 1996; Ratcliff and McKoon, 2008; Forstmann et al., 2016). However, it is unclear whether the brain uses a similar mechanism for making value-based decisions, where the properties of the options being decided between are entirely internal, as opposed to sensory decision-making where those properties are external and directly observable. Some work suggests that, at least behaviorally, reaction times during value-based decision making tasks can be explained using drift-diffusion models (Krajbich and Rangel, 2011).

However, an alternative possibility is that there is no slow integration process, but rather the brain considers options serially and discretely, and makes a decision without evidence accumulation per se. Recent work indicates that, at least in the orbitofrontal cortex, value signals switch back and forth suddenly, suggesting that options are being considered serially during value-based decisions (Rich and Wallis, 2016; Wallis, 2018). Yet a third possibility is that both of these mechanisms occur simultaneously: certain brain areas could consider and evaluate options simultaneously, while others accumulate evidence for each decision and trigger the corresponding action to be taken when some decision threshold is reached.

Here, we investigated if and how the encoding of reward in prefrontal cortex is affected by representations in hippocampus. We performed Bayesian decoding of reward in dmPFC and also Bayesian decoding of position from activity in CA1, while rats deliberated at the choice point of the contingency switch task. But does dmPFC even encode whether animals received reward or not? We performed Bayesian decoding on ensemble activity in dmPFC as rats entered the reward zone, and decoded

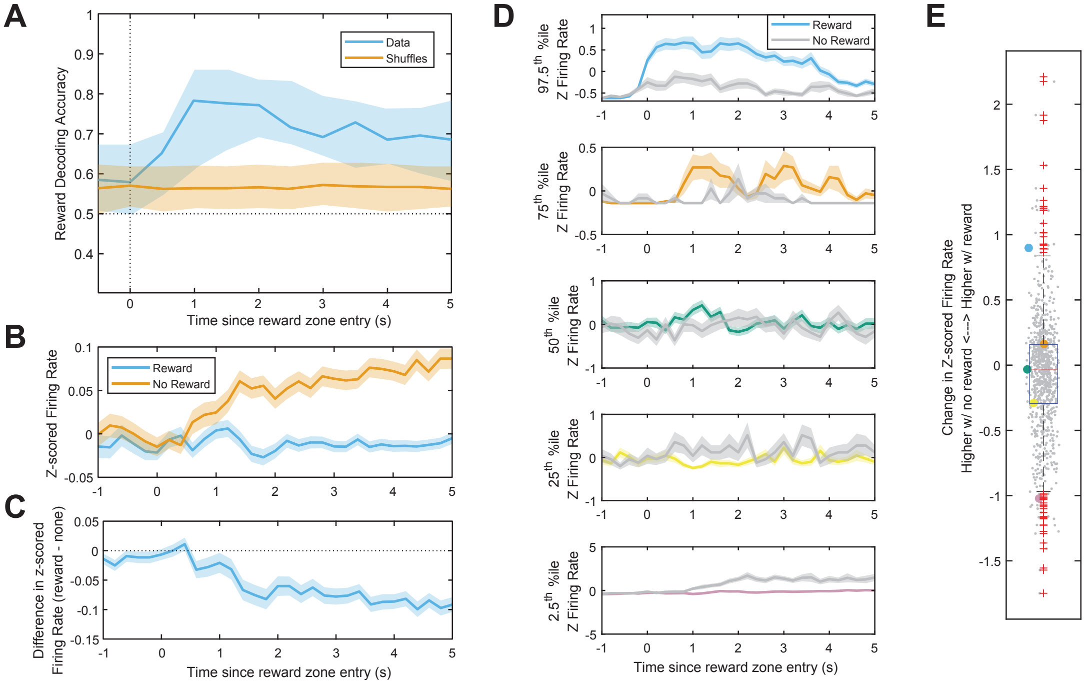

Figure 5.5: Reward representation in dmPFC. (A) Accuracy of Bayesian decoding of reward (vs the lack thereof) from dmPFC ensembles compared to decoding accuracy of Bayesian decoding performed on shuffled firing rates. Lines show the median and shaded areas show the 1σ percentile ($\sim 68.3\%$) across $N = 40$ sessions. (B) z-scored firing rates of individual units following reward zone entry, split by whether reward was received on that lap or not. (C) Difference in z-scored firing rates between reward and lack thereof. Panels B and C show the mean ± standard error across $N = 824$ units. (D) z-scored firing rates upon reward zone entry for individual example cells. Color of the line in each panel corresponds to the colored dots in (E). (E) Difference in z-scored firing rates between reward and lack thereof for all units. Greater values indicate greater firing rates following reward than following a lack of reward, while lesser values indicate a lesser firing rate following reward than following a lack of reward. Note that the median is only slightly less than 0 (the averaging effect seen in panel B and C), but that this difference is small compared to the spread of the entire distribution.

whether rats received reward on that lap or whether there was a lack of reward. The task included an audio cue as to whether a choice was correct or incorrect upon reward zone entry, and so rats had multiple sources of information as to whether their choice was correct or incorrect (both the audio cue and the presence or absence of food reward at the feeder site). The accuracy of reward decoding was significantly above the accuracy of Bayesian decoding performed on shuffled firing rates between around 1-3 seconds after reward zone entry (Figure 5.5A). This indicates that dmPFC does in fact carry information about reward, at least at the time of reward delivery.

How was this reward information represented? Firing rates were, on average, significantly higher after an incorrect choice than after reward was delivered (Figure 5.5B,C). However, to say that firing rates decreased following reward receipt would be a mischaracterization of the data, as the effect was mostly due to averaging. While cells showed a decreased firing rate upon reward on average, there was a wide distribution of responses to reward across individual units. Some cells' firing rates increased in response to reward while other cells had greater firing rates after incorrect choices (Figure 5.5E, and see Figure 5.5D for examples of cells across the distribution). This indicates that the reward encoding we observed in dmPFC was not simply due to "reward cells" or cells which decreased their firing rate in the same direction upon reward receipt, but rather that the representation of reward was due to encoding across the entire ensemble. This is consistent with previous work which finds that single units in prefrontal areas tend to have highly mixed selectivity.

Did prefrontal representations of reward change in response to the encoding of non-local information in hippocampus? We performed Bayesian decoding of reward from ensembles in dmPFC while simultaneously decoding location from ensembles in HPC. The HPC decoder was trained on location and neural activity across the entire session, with 100-fold cross validation, using time bins corresponding to hippocampal theta cycles. The decoder used for dmPFC was trained to decode reward vs the lack thereof, using only epochs 1-3s after reward zone entry, which was the time during which we found dmPFC to be most reliably encoding reward (Figure 5.5A), and evaluated while rats were passing through the choice point. The time bins used for dmPFC decoding were the same as those used for hippocampal decoding (theta cycles of the hippocampal LFP). We split hippocampal theta cycles into those where hippocampus was representing "local" spatial information and those where hippocampus was representing non-local, reward zone information. "Local" theta cycles were those where the decoded spatial posterior distribution had more posterior probability in the choice point than anywhere else on the maze. In contrast, the "nonlocal" theta cycles were those during which the reward zone had more posterior probability than the rest of the maze combined (so, they were not just non-local in a broad sense, but specifically those theta cycles during which hippocampus represented the reward zone most strongly).

We found that, on average, during theta cycles during which HPC was representing the reward zone, there was not a significantly different amount of reward encoding in dmPFC (Figure 5.6A). However, there was a significant change in the reward representation. Reward encoding in dmPFC significantly increased during theta cycles where HPC represented the reward zone ($p = 0.0021$, two-sided Wilcoxon signed rank test comparing the change in decoded reward probabilities from the previous theta cycle, $N = 40$ sessions). In contrast, dmPFC reward encoding did not change during theta cycles where HPC represented only local information ($p = 0.85$, two-sided Wilcoxon signed rank test, $N = 40$ sessions).

To determine the temporal specificity of this effect, we performed a lag analysis similar to that used in Figures 5.4C and D. We looked at the change in reward representation in dmPFC as a function of the number of theta cycles by which the

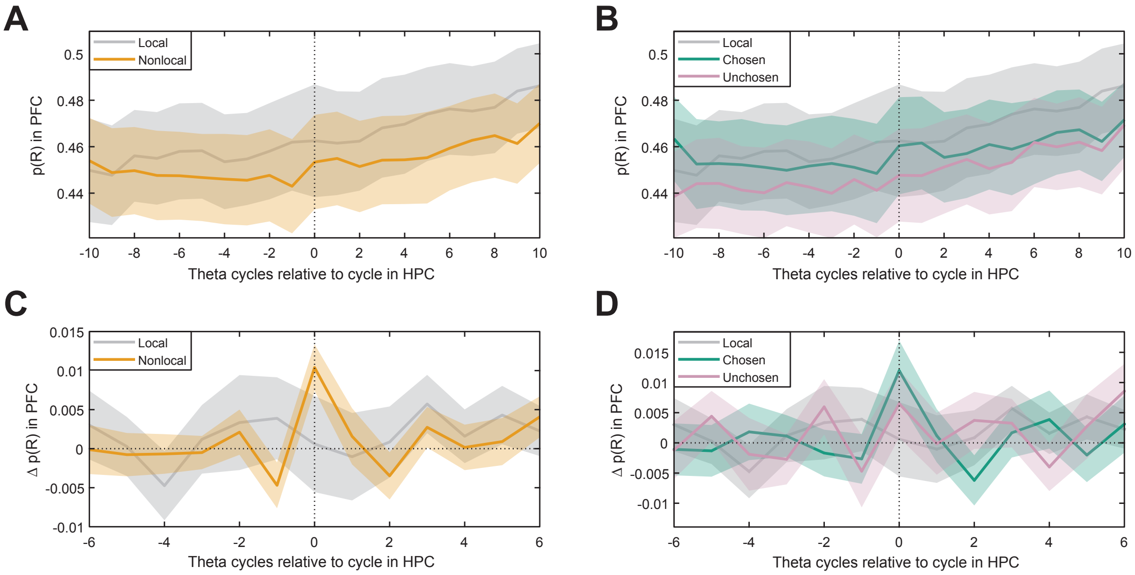

Figure 5.6: Reward encoding in dmPFC and its dependence on goal encoding in HPC. (A) Decoded reward probability from dmPFC as a function of hippocampal theta cycles since the reference theta cycle, split by whether hippocampal ensembles represented the goal location or the current location during the reference theta cycle. (B) Same as in panel A, but theta cycles where HPC represented non-local information have been further split by whether HPC represented the goal location the rat ultimately chose on that lap ("Chosen") or the opposite goal location ("Unchosen"). (C) Change in the decoded reward probability as a function of lag (the derivative of the reward signal in panel A). (D) Change in the decoded reward probability as a function of lag (the derivative of the reward signal shown in panel B). Shown in all panels is the mean ± standard error, $N = 40$ sessions.

reference theta cycle in HPC led or lagged the decoding time bin used for reward decoding from dmPFC. Unlike the results in the previous section (Figures 5.4C and D), we found that the change in reward representation in dmPFC was very tightly locked to the theta cycle during which HPC represented the reward zone (Figure 5.6C). The only offset during which the reward encoding in dmPFC significantly increased was when the same time bin was used for reward decoding from dmPFC and location decoding from HPC (during which HPC represented the reward zone). Theta cycles during which HPC represented only local information showed no systematic change in reward encoding in dmPFC, at any lag (Figure 5.6C). Although on average the reward encoding in dmPFC increased by a greater amount on theta cycles during which HPC represented specifically the reward zone on the side which the rat ended up choosing on that lap, this was not significantly different between theta cycles during which HPC represented the chosen vs the unchosen side (Figure 5.6D). Taken together, these results suggest that hippocampal spatial representations may have a fast, within-theta-cycle, effect on reward encoding in dmPFC.

Chapter 6

Discussion

We investigated choice strategies of rats on a two-step task and found that rats' choices could be explained by a mix of model-based and model-free decision making. Behavioral markers of deliberation such as vicarious trial and error were related to the novelty of choice sequences on the task, while behavioral markers of habit like path stereotypy increased with extended sequences of repeated choices. Vicarious trial and error was correlated between the two choice points, suggesting that rats may enter deliberative modes over the course of entire trials on multi-stage choice tasks. However, we found that our spatial version of the two-step task was overly difficult for rats to learn and was insufficient for providing enough data per session to fit reinforcement learning models.

Therefore, we designed a new variant of the contingency switching task, to enable the study of alternations between deliberation and habitual modes during decision-making. Vicarious trial and error was related to rats' uncertainty as to the task contingencies on this new task. We also found that both CA1 and dmPFC encoded the contingencies while simultaneously representing other information which changed over time. Ensembles in dmPFC began to represent the new contingency before rats exhibited behavioral changes, while ensembles in HPC began to represent the new contingency later only when rats updated their choices to be consistent with the new contingency.

Lastly, we examined the relationship between hippocampal and prefrontal representations on faster timescales, and found that they were related in complex ways. Theta oscillations in the two areas were coherent, especially at choice points on the contingency switching task, and especially following errors. While HPC and dmPFC spatial representations were correlated in that they represented either local information or prospective information together, the two areas did not appear to always represent identical prospective information at the same time. Furthermore, activity in prefrontal predicted whether hippocampus was representing local or prospective information, but this relationship occurred across a very broad timescale, on the order of seconds. On the other hand, representations of goal location in hippocampus appeared to have very temporally specific effects on reward encoding in dmPFC, across timescales on the order of a single theta cycle.

To further study the neural underpinnings of model-based and model-free influences on decision making, future work will have to develop and validate a version of the two-step task which works well for rodents. Our full version of the task had some major drawbacks, including the speed of the reward drift (which was too slow) and the low number of trials rats were able to run on the task in a single session. Miller et al. (2017) have developed a different version of the two-step task for rats which depends on a simplified task structure, making it easier for rats to learn. Their version of the task also employs fast, block-like switches in reward contingencies. Instead of slowly drifting reward probabilities, reward probabilities on their version of the task switch suddenly back and forth from 80%/20% to 20%/80% every 30 laps or so. This creates a large, sudden change in the valuations of the model-based and model-free systems:

the model-based system is able to update the valuation after the switch quickly, while the model-free system takes longer, and because the difference in the valuations is so large, the action probabilities of the two algorithms is large, allowing model fits to more accurately captures differences in when each algorithm is likely to be driving behavior. However, their version of the task uses a simplified task structure, and is non-spatial, which prevents using that version of the task to investigate hippocampal spatial representations. Furthermore, both our version of the two-step task and theirs use proxies for cost. We use the time delay to food delivery as the cost, while they use the probability of food delivery at all as the cost. This further complicates the modeling process, as it is not known exactly how delay or probability correspond to the amount of food reward in terms of valuation.

There are also improvements that can be made to the models being fit to animal behavior on the two-step task. In the brain, the model-free and model-based systems are thought to depend on different subsystems. Specifically, the model-free system is thought to have a much slower learning rate (thus giving rise to the habitual or procedural behavior which is inflexible once learned). However, in research investigating algorithms where the model-free and model-based algorithms are combined into a single agent, the two algorithms share learning rates (such as in the constant-weight algorithm in section 2.4.3, also used in much if not all of the human work (Gläscher et al., 2010; Daw et al., 2011; Gillan et al., 2011; Wunderlich et al., 2012; Otto et al., 2013b,a; Eppinger et al., 2013; Skatova et al., 2013; Schad et al., 2014; Gillan et al., 2014; Sebold et al., 2014; Otto et al., 2015; Gillan et al., 2015; Voon et al., 2015; Deserno et al., 2015; Radenbach et al., 2015; Sharp et al., 2015; Doll et al., 2016; Decker et al., 2016), or the uncertainty-based models described in section 2.5). While parameter-efficient, and thus leading to better model comparison scores, this is neurophysiologically unrealistic. Behavioral work strongly suggests that the two systems have drastically different learning rates: the deliberative system is thought to have a very fast learning rate, while the procedural system is hypothesized to have a comparatively slow learning rate. Future work should investigate using models where the two systems have their own learning rates, allowing for better and more realistic explanations of the algorithms driving animals' choices. However, the two systems are not completely independent. It is very likely that the two systems do share some but not all information, which further complicates the modeling.

Another avenue for future research is to investigate how and why slow representational changes over time occur, and whether these changes are due only to random drift, or systematic and predictable. One simple explanation for the observation of representational drift is mixed selectivity, where the brain area in question represents not only the otherwise stable information (in our study, information about contingency), but simultaneously represents information about other factors which are changing over time. This changing information could be any combination of many different factors, but some likely candidates include the representation of motivational state, hunger or thirst, arousal, or awareness. However, some theories suggest that a change in encoding over time is required for fast acquisition of memories in labile states which solidify into stable representations to allow for the reliable storage of those memories (Benna and Fusi, 2016), or perhaps due to the transfer of information from less stable to more stable brain regions or sub-networks (Roxin and Fusi, 2013). This change from labile to stable representation may be causing changes in encoding differences which, having recorded only from a subset of the neurons involved, we observe as representational drift. Other theories suggest that the brain explicitly encodes temporal context so that memories which occurred closer in time can be more easily co-retrieved (Mensink and Raaijmakers, 1988; Howard and Kahana, 2002). In these theories, representational drift over time is explicitly built in to information encoding, such that information encoded at similar time point share similar encoding. In theory, in the presence of associative network dynamics, this time-based encoding would facilitate the retrieval of memories which are closer in time. However, this theory requires more experimental evidence, and from our results we cannot determine whether drift is occurring due to dynamics similar to those described by this theory of temporal context, due to the transfer of memories from labile to stable states, or due simply to the encoding of other unrelated information which is changing over time.

The issue of representational drift also raises the question of how stable representations are even able to occur in the presence of representational drift over time. One possibility is that specific stable representations are encoded by a certain subset of neurons, while other neurons encode information which changes over time, giving rise to observation of ensemble drift over time. However, this explanation seems unlikely, especially in prefrontal cortex, where we know that cell activity exhibits highly mixed selectivity, and single neurons are not often tuned specifically to single concepts. It is also likely that this differs from brain area to brain area. For example, in hippocampus it is much more likely that certain ensembles of neurons represent information stably while others encode information that is changing over time. Although, some work suggests it is truly a phenomenon which occurs across the entire ensemble (Mankin et al., 2012). If so, future research will have to work out how stable representations are even possible in the face of representational drift – or at least how downstream structures parse out stable information from changing inputs.

If it is the case that representational drift occurs across entire ensembles while stable representations are simultaneously maintained, analytical tools to separate these two factors will also have to be developed. Here, we used a linear discriminant analysis based approach to separate contributions of stable representations (of contingency) and drift over time. This method is preferable to training a probabilistic classifier to distinguish between Y_1 and X, and then evaluating on Y_2, because it allows us to see the effect of drift. Using a regular classifier and comparing class probabilities cannot distinguish between the situation where neural activity during Y_2 is not different from the neural activity during X, and the situation where it is more different from Y_1 than is the neural activity during X. Using the LDA analysis, we are able to distinguish these two scenarios because the Y_2 projections are more highly separated from Y_1 projections in the case of drift (Figure 4.3C), whereas in the case of contingency representations the Y_2 projections are more similar to the Y_1 projections than are the X projections (Figures 4.3B,E,F).

While advantageous compared to previous methods, our linear-discriminant-based approach is obviously not well-suited to extracting nonlinear information from ensemble activity, nor does it explicitly model the simultaneous representation of stable information and a non-stable representational drift over time. Perhaps a more principled approach would be to use a model which explicitly accounts for these two contributions to neural activity – some type of mixture model which includes a fast, stable component, but also a slow, drifting component. This could be done, for example, by a model which uses a hidden Markov model to model the stable components which change quickly, in combination with a latent state space model which models the slowly changing representational components. However, combining hidden Markov models and state space models in this way can often lead to underdetermined models (because either the hidden Markov component could change quickly while the state space component changes slowly, or vice-versa, either situation being equally as likely given certain parameter values). New mathematical approaches will have to be developed which address this difficulty, in order to model the representation of stable information with a simultaneous representational drift over time.

The work done here analyzed differences in the timing of switches in contingency representation between dorsomedial prefrontal cortex and hippocampus. Our interpretation was that information about contingency from higher cortical areas such as prefrontal cortex takes time to influence the more intrinsically stable representations in hippocampus. However, our experiments are not able to determine a causal influence of contingency information in prefrontal areas on hippocampal representations. To truly determine whether representations about contingency in hippocampal areas arise due to inputs from prefrontal areas, future work will have to perform inactivation studies which investigate the causal effect of prefrontal inputs to the hippocampus. Some work inactivating the prefrontal cortex indeed finds disruptions in prospective representations in the hippocampus during deliberation (Schmidt et al., 2019), but similar experiments have yet to be performed which achieve directional selectivity on fast timescales. The situation is further complicated by the distributed nature of the deliberative network – for example inhibiting prefrontal afferents to the hippocampus (or to the nucleus reuniens or other intermediary structures) does not preclude that information arriving in the hippocampus via other routes, or even arising in different brain structures (such as orbitofrontal cortex). So, future work will need to combine temporally-specific inactivation methods (such as optogenetics) with analyses of the timing of representational changes in the relevant structures in order to disambiguate how different forms of contextual information reach the hippocampus from higher cortical areas.

Similarly, our analyses were not able to parse out the directionality of information flow during non-local representations. Here we found that non-local representations in the dorsomedial prefrontal cortex and hippocampus were correlated, and that there may have been a relationship between non-local representation in hippocampus and reward encoding in prefrontal cortex. However, without causal manipulations, we were unable to determine the directionality of these relationships. It could be that non-local representations in hippocampus cause corresponding reward encoding to occur in prefrontal cortex, perhaps due to the activation of intermediate reward-related structures like the ventral striatum (van der Meer and Redish, 2010; van der Meer et al., 2012). Alternatively, it could be that representations of goal-related information in prefrontal cortex, in concert with information about motivational state in other brain areas, causes the hippocampus to represent potential paths toward the goal being represented in prefrontal cortex. Finally, because of the bidirectional nature of connectivity between HPC and dmPFC, the directionality of information flow may be complex, or even not clearly directional. Though some work suggests that the flow of information between prefrontal and hippocampal areas varies across time and depends on task demands (Jones and Wilson, 2005b; Bähner et al., 2015; Shin and Jadhav, 2016). To causally determine how prospective and goal-relevant information flows between prefrontal and hippocampal areas, future work will need to employ temporally-specific and perhaps even projection-specific causal manipulation methods.

Finally, perhaps instead of trying to understand neural systems in terms of old reinforcement learning models, the field should attempt to formulate more neurophysiologically-driven theories of action selection, along the lines of theoretical work focusing on the habitual system, such as Frank (2011). Certain assumptions of the model-based and model-free algorithms being used to explain animal behavior simply do not align with knowledge of how the brain stores or recalls information. For example, the model-free algorithm uses a "lookup table" for action selection: it stores the expected reward (the Q-value) of every possible action in *every* state, and at decision time, "looks up" the action with the highest expected reward (or, technically, performs a softmax over action values). However, this is clearly more of an analogy than realistic characterization of how the brain works. Instead, it is though that the dorsolateral striatum simply associates state information (coming from sensory association cortical areas) with action information, encoded in its outputs to thalamus and later basal ganglia structures, and that this association is trained by the release of dopamine from the VTA/SNc. Instead of focusing on matching the deliberative system to model-based algorithms, future research may be better served by investigating more neurophysiologically-driven theories of how the neural deliberative system functions.

It remains an open question how the deliberative system evaluates its internal model, what causes this evaluation to occur, and even whether this evaluation is as discrete and explicit as theories suggest. Certainly current work suggests that some internal simulation process does occur (Johnson and Redish, 2007; Hassabis and Maguire, 2009; Wang et al., 2015), and further work suggests that the outcomes of these internal simulations are internally evaluated (van der Meer and Redish, 2010; van der Meer et al., 2012; Rich and Wallis, 2016), but it is unknown how these events are instigated. It could be that higher-level brain areas such as parts of the prefrontal cortex identify a need for the deliberative system to be engaged (perhaps because simpler and faster systems like the procedural system have not generated candidate actions). In this thesis, we have assumed that the deliberative or model-based system works in this way by actively instigating an explicit, discrete simulation of potential futures, particularly in chapters 2 and 3, where we simulated model-based algorithms. However, it is very possible that this process is more passive in nature. To make progress, future work may be well-served by moving away from attempting to match the neurophysiology to existing reinforcement learning algorithms, and instead begin to shift towards developing neurally-inspired theories to model the inner workings of decision systems.

Bibliography

C. Adams and A. Dickinson. Actions and habits: Variations in associative representations during instrumental learning. In *Information processing in animals: Memory mechanisms*, pages 143–165. Erlbaum Hillsdale, NJ, 1981.

T. Akam, P. Dayan, and R. Costa. Multi-step decision tasks for dissociating model-based and model-free learning in rodents. Presentation at the Computational and Systems Neuroscience Conference (Cosyne), Salt Lake City, UT, 2013.

T. Akam, R. Costa, and P. Dayan. Simple plans or sophisticated habits? state, transition and learning interactions in the two-step task. *PLoS computational biology*, 11(12):e1004648, 2015. doi:10.1371/journal.pcbi.1004648.

Amir S Bahar, Prasad R Shirvalkar, and Matthew L Shapiro. Memory-guided learning: Ca1 and ca3 neuronal ensembles differentially encode the commonalities and differences between situations. *Journal of Neuroscience*, 31(34):12270–12281, 2011. doi:10.1523/JNEUROSCI.1671-11.2011.

F. Bähner, C. Demanuele, J. Schweiger, M.F. Gerchen, V. Zamoscik, K. Ueltzhöffer, T. Hahn, P. Meyer, H. Flor, D. Durstewitz, et al. Hippocampal–dorsolateral prefrontal coupling as a species-conserved cognitive mechanism: A human translational imaging study. *Neuropsychopharmacology*, 40(7):1674, 2015. doi:10.1038/npp.2015.13.

Bernard W Balleine and Anthony Dickinson. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37(4-5):407–419, 1998. doi:10.1016/S0028-3908(98)00033-1.

B.W. Balleine, N.D. Daw, and J.P. O'Doherty. Multiple forms of value learning and the function of dopamine. In *Neuroeconomics: decision making and the brain*, pages 367–385. Academic Press Waltham, MA, 2008.

Gareth RI Barker, Paul J Banks, Hannah Scott, G Scott Ralph, Kyriacos A Mitrophanous, Liang-Fong Wong, Zafar I Bashir, James B Uney, and E Clea Warburton. Separate elements of episodic memory subserved by distinct hippocampal–prefrontal connections. *Nature neuroscience*, 20(2):242, 2017. doi:10.1038/nn.4472.

U.R. Beierholm, C. Anen, S. Quartz, and P. Bossaerts. Separate encoding of model-based and model-free valuations in the human brain. *Neuroimage*, 58(3):955–962, 2011. doi:10.1016/j.neuroimage.2011.06.071.

Karim Benchenane, Adrien Peyrache, Mehdi Khamassi, Patrick L Tierney, Yves Gioanni, Francesco P Battaglia, and Sidney I Wiener. Coherent theta oscillations and reorganization of spike timing in the hippocampal-prefrontal network upon learning. *Neuron*, 66(6):921–936, 2010. doi:10.1016/j.neuron.2010.05.013.

Marcus K Benna and Stefano Fusi. Computational principles of synaptic memory consolidation. *Nature neuroscience*, 19(12):1697, 2016. doi:10.1038/nn.4401.

Anna Blumenthal, Adam Steiner, Kelsey Seeland, and A. David Redish. Effects of pharmacological manipulations of nmda-receptors on deliberation in the multiple-t task. *Neurobiology of Learning and Memory*, 95(3):376 – 384, 2011. doi:10.1016/j.nlm.2011.01.011.

Yannick-André Breton, Kelsey D Seeland, and A David Redish. Aging impairs deliberation and behavioral flexibility in inter-temporal choice. *Frontiers in aging neuroscience*, 7:41, 2015. doi:10.3389/fnagi.2015.00041.

Scott L Brincat and Earl K Miller. Frequency-specific hippocampal-prefrontal interactions during associative learning. *Nature neuroscience*, 18(4):576, 2015. doi:10.1038/nn.3954.

TI Brown, VA Carr, KF LaRocque, SE Favila, AM Gordon, B Bowles, JN Bailenson, and AD Wagner. Prospective representation of navigational goals in the human hippocampus. *Science*, 352(6291):1323–1326, 2016.

P. Calabresi, B. Picconi, A. Tozzi, and M. Di Filippo. Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends in neurosciences*, 30(5):211–219, 2007. doi:10.1016/j.tins.2007.03.001.

B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32, 2017. ISSN 1548-7660. doi:10.18637/jss.v076.i01.

Jean-Christophe Cassel, Anne Pereira De Vasconcelos, Michaël Loureiro, Thibault Cholvin, John C Dalrymple-Alford, and Robert P Vertes. The reuniens and rhomboid nuclei: neuroanatomy, electrophysiological characteristics and behavioral implications. *Progress in neurobiology*, 111:34–52, 2013. doi:10.1016/j.pneurobio.2013.08.006.

N.J. Cohen and H.B. Eichenbaum. *Memory, amnesia, and hippocampal function.* MIT Press, 1993.

Laura Lee Colgin. Oscillations and hippocampal–prefrontal synchrony. *Current opinion in neurobiology*, 21(3):467–474, 2011. doi:10.1016/j.conb.2011.04.006.

A. G. E. Collins and M. J. Frank. Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological review*, 120(1):190, 2013. doi:10.1037/a0030852.

Stephen L Cowen and Bruce L McNaughton. Selective delay activity in the medial prefrontal cortex of the rat: contribution of sensorimotor information and contingency. *Journal of neurophysiology*, 98(1):303–316, 2007. doi:10.1152/jn.00150.2007.

Jeffrey W Dalley, Rudolf N Cardinal, and Trevor W Robbins. Prefrontal executive and cognitive functions in rodents: neural and neurochemical substrates. *Neuroscience & Biobehavioral Reviews*, 28(7):771–784, 2004. doi:10.1016/j.neubiorev.2004.09.006.

N.D. Daw and P. Dayan. The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369 (1655), 2014. ISSN 0962-8436. doi:10.1098/rstb.2013.0478.

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

N.D. Daw, S.J. Gershman, B. Seymour, P. Dayan, and R.J. Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011. doi:10.1016/j.neuron.2011.02.027.

P. Dayan and B.W. Balleine. Reward, motivation, and reinforcement learning. *Neuron*, 36(2):285–298, 2002. doi:10.1016/S0896-6273(02)00963-7.

Emanuela De Falco, Lei An, Ninglei Sun, Andrew J Roebuck, Quentin Greba, Christopher C Lapish, and John G Howland. The rat medial prefrontal cortex exhibits flexible neural activity states during the performance of an odor span task. *eNeuro*, 6(2), 2019. doi:10.1523/ENEURO.0424-18.2019.

J.H. Decker, A.R. Otto, N.D. Daw, and C.A. Hartley. From creatures of habit to goal-directed learners tracking the developmental emergence of model-based reinforcement learning. *Psychological science*, 27(6):848–858, 2016. doi:10.1177/0956797616639301.

B Delatour and MP Witter. Projections from the parahippocampal region to the prefrontal cortex in the rat: evidence of multiple pathways. *European Journal of Neuroscience*, 15(8):1400–1407, 2002. doi:10.1046/j.1460-9568.2002.01973.x.

Benoît Delatour and Pascale Gisquest-Verrier. Lesions of the prelimbic–infralimbic cortices in rats do not disrupt response selection processes but induce delay-dependent deficits: evidence for a role in working memory? *Behavioral neuroscience*, 113(5):941, 1999. doi:10.1037/0735-7044.113.5.941.

L. Deserno, Q.J.M. Huys, R. Boehme, R. Buchert, H. Heinze, A.A. Grace, R.J. Dolan, A. Heinz, and F. Schlagenhauf. Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences*, 112(5):1595–1600, 2015. doi:10.1073/pnas.1417219112.

A. Dezfouli and B.W. Balleine. Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, 35(7):1036–1051, 2012. doi:10.1111/j.1460-9568.2012.08050.x.

A. Dezfouli and B.W. Balleine. Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLoS computational biology*, 9(12):e1003364, 2013. doi:10.1371/journal.pcbi.1003364.

A. Dezfouli, N.W. Lingawi, and B.W. Balleine. Habits as action sequences: hierarchical action control and changes in outcome value. *Phil. Trans. R. Soc. B*, 369 (1655):20130482, 2014. doi:10.1098/rstb.2013.0482.

Gonzalo Viana Di Prisco and Robert P Vertes. Excitatory actions of the ventral midline thalamus (rhomboid/reuniens) on the medial prefrontal cortex in the rat. *Synapse*, 60(1):45–55, 2006. doi:10.1002/syn.20271.

R.J. Dolan and P. Dayan. Goals and habits in the brain. *Neuron*, 80(2):312–325, 2013. doi:10.1016/j.neuron.2013.09.007.

B.B. Doll, D.A. Simon, and N.D. Daw. The ubiquity of model-based reinforcement learning. *Current opinion in neurobiology*, 22(6):1075–1081, 2012. doi:10.1016/j.conb.2012.08.003.

B.B. Doll, K.D. Duncan, D.A. Simon, D. Shohamy, and N.D. Daw. Model-based choices involve prospective neural activity. *Nature Neuroscience*, 18(5):767–772, 2015. doi:10.1126/science.aaf0784.

B.B. Doll, K.G. Bath, N.D. Daw, and M.J. Frank. Variability in dopamine genes dissociates model-based and model-free reinforcement learning. *The Journal of Neuroscience*, 36(4):1211–1222, 2016. doi:10.1523/JNEUROSCI.1901-15.2016.

Margriet J Dolleman-van der Weel, Richard GM Morris, and Menno P Witter. Neurotoxic lesions of the thalamic reuniens or mediodorsal nucleus in rats affect non-mnemonic aspects of watermaze learning. *Brain structure and function*, 213(3): 329–342, 2009. doi:10.1007/s00429-008-0200-6.

Margriet J Dolleman-van der Weel, Amy L Griffin, Hiroshi T Ito, Matthew L Shapiro, Menno P Witter, Robert P Vertes, and Timothy A Allen. The nucleus reuniens of the thalamus sits at the nexus of a hippocampus and medial prefrontal cortex circuit enabling memory and behavior. *Learning & Memory*, 26(7):191–205, 2019. doi:10.1101/lm.048389.118.

MJ Dolleman-Van der Weel, FH Lopes Da Silva, and Menno P Witter. Interaction of nucleus reuniens and entorhinal cortex projections in hippocampal field ca1 of the rat. *Brain Structure and Function*, 222(5):2421–2438, 2017. doi:10.1007/s00429-016-1350-6.

Daniel Durstewitz, Nicole M Vittoz, Stan B Floresco, and Jeremy K Seamans. Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron*, 66(3):438–448, 2010. doi:10.1016/j.neuron.2010.03.029.

Dominic M Dwyer, Michael J Dunn, Sarah EV Rhodes, and A Simon Killcross. Lesions of the prelimbic prefrontal cortex prevent response conflict produced by action–outcome associations. *Quarterly Journal of Experimental Psychology*, 63 (3):417–424, 2010. doi:10.1080/17470210903411049.

Howard Eichenbaum. Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*, 18(9):547, 2017. doi:10.1038/nrn.2017.74.

B. Eppinger, M. Walter, H.R. Heekeren, and S.C. Li. Of goals and habits: age-related and individual differences in goal-directed decision-making. *Frontiers in neuroscience*, 7:253, 2013. doi:10.3389/fnins.2013.00253.

David R Euston, Aaron J Gruber, and Bruce L McNaughton. The role of medial prefrontal cortex in memory and decision making. *Neuron*, 76(6):1057–1070, 2012. doi:10.1016/j.neuron.2012.12.002.

Janina Ferbinteanu and Matthew L Shapiro. Prospective and retrospective memory coding in the hippocampus. *Neuron*, 40(6):1227–1239, 2003. doi:10.1016/S0896-6273(03)00752-9.

Janina Ferbinteanu, Prasad Shirvalkar, and Matthew L Shapiro. Memory modulates journey-dependent coding in the rat hippocampus. *Journal of Neuroscience*, 31 (25):9135–9146, 2011. doi:10.1523/JNEUROSCI.1241-11.2011.

F Ferino, AM Thierry, and J Glowinski. Anatomical and electrophysiological evidence for a direct projection from ammon's horn to the medial prefrontal cortex in the rat. *Experimental brain research*, 65(2):421–426, 1987. doi:10.1007/bf00236315.

Stan B Floresco and Anthony A Grace. Gating of hippocampal-evoked activity in prefrontal cortical neurons by inputs from the mediodorsal thalamus and ventral tegmental area. *Journal of Neuroscience*, 23(9):3930–3943, 2003. doi:10.1523/JNEUROSCI.23-09-03930.2003.

Stan B Floresco, Jeremy K Seamans, and Anthony G Phillips. Selective roles for hippocampal, prefrontal cortical, and ventral striatal circuits in radial-arm maze tasks with or without a delay. *Journal of Neuroscience*, 17(5):1880–1890, 1997. doi:10.1523/JNEUROSCI.17-05-01880.1997.

Stan B Floresco, Annie E Block, and TL Maric. Inactivation of the medial prefrontal cortex of the rat impairs strategy set-shifting, but not reversal learning, using a novel, automated procedure. *Behavioural brain research*, 190(1):85–96, 2008. doi:10.1016/j.bbr.2008.02.008.

Birte U Forstmann, Roger Ratcliff, and E-J Wagenmakers. Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual review of psychology*, 67:641–666, 2016. doi:10.1146/annurev-psych-122414-033645.

M.J. Frank. Computational models of motivated action selection in corticostriatal circuits. *Current opinion in neurobiology*, 21(3):381–386, 2011. doi:10.1016/j.conb.2011.02.013.

M.C. Fuhs and D.S. Touretzky. Context learning in the rodent hippocampus. *Neural Computation*, 19(12):3173–3215, 2007. doi:10.1162/neco.2007.19.12.3173.

Shigeyoshi Fujisawa and György Buzsáki. A 4 hz oscillation adaptively synchronizes prefrontal, vta, and hippocampal activities. *Neuron*, 72(1):153–165, 2011. doi:10.1016/j.neuron.2011.08.018.

Charles R Gallistel, Stephen Fairhurst, and Peter Balsam. The learning curve: implications of a quantitative analysis. *Proceedings of the National Academy of Sciences*, 101(36):13124–13131, 2004. doi:10.1073/pnas.0404965101.

R.S. Gardner, M.R. Uttaro, S.E. Fleming, D.F. Suarez, G.A. Ascoli, and T.C. Dumas. A secondary working memory challenge preserves primary place strategies despite overtraining. *Learning & Memory*, 20(11):648–656, 2013. doi:10.1101/lm.031336.113.

Samuel J Gershman, Kenneth A Norman, and Yael Niv. Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, 5:43–50, 2015. doi:10.1016/j.cobeha.2015.07.007.

C.M. Gillan, M. Papmeyer, S. Morein-Zamir, B.J. Sahakian, N.A. Fineberg, T.W. Robbins, and S. de Wit. Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *American Journal of Psychiatry*, 168(7):718–726, 2011. doi:10.1176/appi.ajp.2011.10071062.

C.M. Gillan, S. Morein-Zamir, M. Kaser, N.A. Fineberg, A. Sule, B.J. Sahakian, R.N. Cardinal, and T.W. Robbins. Counterfactual processing of economic action-outcome alternatives in obsessive-compulsive disorder: further evidence of impaired goal-directed behavior. *Biological psychiatry*, 75(8):639–646, 2014. doi:10.1016/j.biopsych.2013.01.018.

C.M. Gillan, A.R. Otto, E.A. Phelps, and N.D. Daw. Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience*, 15(3): 523–536, 2015. doi:10.3758/s13415-015-0347-6.

J. Gläscher, N.D. Daw, P. Dayan, and J.P. O'Doherty. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–595, 2010. doi:10.1016/j.neuron.2010.04.016.

Joshua A Gordon. Oscillations and hippocampal–prefrontal synchrony. *Current opinion in neurobiology*, 21(3):486–491, 2011. doi:10.1016/j.conb.2011.02.012.

Amy L Griffin. Role of the thalamic nucleus reuniens in mediating interactions between the hippocampus and medial prefrontal cortex during spatial working memory. *Frontiers in systems neuroscience*, 9:29, 2015. doi:10.3389/fnsys.2015.00029.

Amy L Griffin, Howard Eichenbaum, and Michael E Hasselmo. Spatial representations of hippocampal ca1 neurons are modulated by behavioral context in a hippocampus-dependent memory task. *Journal of Neuroscience*, 27(9):2416–2423, 2007. doi:10.1523/JNEUROSCI.4083-06.2007.

Kevin G Guise and Matthew L Shapiro. Medial prefrontal cortex reduces memory interference by modifying hippocampal encoding. *Neuron*, 94(1):183–192, 2017. doi:10.1016/j.neuron.2017.03.011.

Anoopum S Gupta, Matthijs AA van der Meer, David S Touretzky, and A David Redish. Hippocampal replay is not a simple function of experience. *Neuron*, 65(5): 695–705, 2010. doi:10.1016/j.neuron.2010.01.034.

Anoopum S Gupta, Matthijs AA van der Meer, David S Touretzky, and A David Redish. Segmentation of spatial experience by hippocampal theta sequences. *Nature neuroscience*, 15(7):1032, 2012. doi:10.1038/nn.3138.

Josephine E Haddon and Simon Killcross. Prefrontal cortex lesions disrupt the contextual control of response conflict. *Journal of Neuroscience*, 26(11):2933–2940, 2006. doi:10.1523/JNEUROSCI.3243-05.2006.

Josephine Elizabeth Haddon and Andrew Simon Killcross. Medial prefrontal cortex lesions abolish contextual control of competing responses. *Journal of the experimental analysis of behavior*, 84(3):485–504, 2005. doi:10.1901/jeab.2005.81-04.

Henry L Hallock, Arick Wang, Crystal L Shaw, and Amy L Griffin. Transient inactivation of the thalamic nucleus reuniens and rhomboid nucleus produces deficits of a working-memory dependent tactile-visual conditional discrimination task. *Behavioral neuroscience*, 127(6):860, 2013. doi:10.1037/a0034653.

Henry L Hallock, Arick Wang, and Amy L Griffin. Ventral midline thalamus is critical for hippocampal–prefrontal synchrony and spatial working memory. *Journal of Neuroscience*, 36(32):8372–8389, 2016. doi:10.1523/JNEUROSCI.0991-16.2016.

Doug P Hanes and Jeffrey D Schall. Neural control of voluntary movement initiation. *Science*, 274(5286):427–430, 1996. doi:10.1126/science.274.5286.427.

Demis Hassabis and Eleanor A Maguire. The construction system of the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1263–1271, 2009. doi:10.1098/rstb.2008.0296.

Michael E Hasselmo and Howard Eichenbaum. Hippocampal mechanisms for the context-dependent retrieval of episodes. *Neural networks*, 18(9):1172–1190, 2005. doi:10.1016/j.neunet.2005.08.007.

B. Hasz and A.D. Redish. Deliberation and procedural automation on a two-step task for rats. *Frontiers in integrative neuroscience*, 12, 2018. doi:10.3389/fnint.2018.00030.

V Hok, E Save, PP Lenck-Santini, and B Poucet. Coding for spatial goals in the prelimbic/infralimbic area of the rat frontal cortex. *Proceedings of the National Academy of Sciences*, 102(12):4602–4607, 2005. doi:10.1073/pnas.0407332102.

Vincent Hok, Ehsan Chah, Etienne Save, and Bruno Poucet. Prefrontal cortex focally modulates hippocampal place cell firing patterns. *Journal of Neuroscience*, 33(8): 3443–3451, 2013. doi:10.1523/JNEUROSCI.3427-12.2013.

Walter B Hoover and Robert P Vertes. Anatomical analysis of afferent projections to the medial prefrontal cortex in the rat. *Brain Structure and Function*, 212(2): 149–179, 2007. doi:10.1007/s00429-007-0150-4.

Walter B Hoover and Robert P Vertes. Collateral projections from nucleus reuniens of thalamus to hippocampus and medial prefrontal cortex in the rat: a single and double retrograde fluorescent labeling study. *Brain Structure and Function*, 217 (2):191–209, 2012. doi:10.1007/s00429-011-0345-6.

Nicole K Horst and Mark Laubach. The role of rat dorsomedial prefrontal cortex in spatial working memory. *Neuroscience*, 164(2):444–456, 2009. doi:10.1016/j.neuroscience.2009.08.004.

Marc W Howard and Michael J Kahana. A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46(3):269–299, 2002. doi:10.1006/jmps.2001.1388.

James M Hyman, Eric A Zilli, Amanda M Paley, and Michael E Hasselmo. Medial prefrontal cortex cells show dynamic modulation with the hippocampal theta rhythm dependent on behavior. *Hippocampus*, 15(6):739–749, 2005. doi:10.1002/hipo.20106.

James M Hyman, Eric A Zilli, Amanda M Paley, and Michael E Hasselmo. Working memory performance correlates with prefrontal-hippocampal theta interactions but not with prefrontal neuron firing rates. *Frontiers in integrative neuroscience*, 4:2, 2010. doi:10.3389/neuro.07.002.2010.

James M Hyman, Liya Ma, Emili Balaguer-Ballester, Daniel Durstewitz, and Jeremy K Seamans. Contextual encoding by ensembles of medial prefrontal cortex neurons. *Proceedings of the National Academy of Sciences*, 109(13):5086–5091, 2012. doi:10.1073/pnas.1114415109.

James Michael Hyman, Michael Erik Hasselmo, and Jeremy Keith Seamans. What is the functional relevance of prefrontal cortex entrainment to hippocampal theta rhythms? *Frontiers in neuroscience*, 5:24, 2011. doi:10.3389/fnins.2011.00024.

Hiroshi T Ito, Sheng-Jia Zhang, Menno P Witter, Edvard I Moser, and May-Britt Moser. A prefrontal–thalamo–hippocampal circuit for goal-directed spatial navigation. *Nature*, 522(7554):50, 2015. doi:10.1038/nature14396.

Hiroshi T Ito, Edvard I Moser, and May-Britt Moser. Supramammillary nucleus modulates spike-time coordination in the prefrontal-thalamo-hippocampal circuit during navigation. *Neuron*, 99(3):576–587, 2018. doi:10.1016/j.neuron.2018.07.021.

Y Yu Jai and Loren M Frank. Hippocampal–cortical interaction in decision making. *Neurobiology of learning and memory*, 117:34–41, 2015. doi:10.1016/j.nlm.2014.02.002.

Thérèse M Jay and Menno P Witter. Distribution of hippocampal ca1 and subicular efferents in the prefrontal cortex of the rat studied by means of anterograde transport of phaseolus vulgaris-leucoagglutinin. *Journal of Comparative Neurology*, 313 (4):574–586, 1991. doi:10.1002/cne.903130404.

M.S. Jog, Y. Kubota, C.I. Connolly, V. Hillegaart, and A.M. Graybiel. Building neural representations of habits. *Science*, 286(5445):1745–1749, 1999. doi:10.1126/science.286.5445.1745.

A. Johnson and A.D. Redish. Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. *The Journal of neuroscience*, 27(45): 12176–12189, 2007. doi:10.1523/JNEUROSCI.3761-07.2007.

Matthew W Jones and Matthew A Wilson. Phase precession of medial prefrontal cortical activity relative to the hippocampal theta rhythm. *Hippocampus*, 15(7): 867–873, 2005a. doi:10.1002/hipo.20119.

Matthew W Jones and Matthew A Wilson. Theta rhythms coordinate hippocampal–prefrontal interactions in a spatial memory task. *PLoS biology*, 3(12):e402, 2005b. doi:10.1371/journal.pbio.0030402.

Min W Jung, Yulin Qin, Bruce L McNaughton, and Carol A Barnes. Firing characteristics of deep layer neurons in prefrontal cortex in rats performing spatial working memory tasks. *Cerebral cortex (New York, NY: 1991)*, 8(5):437–450, 1998. doi:10.1093/cercor/8.5.437.

Daniel Kahneman. *Thinking, fast and slow.* Macmillan, 2011.

Mattias P Karlsson, Dougal GR Tervo, and Alla Y Karpova. Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science*, 338 (6103):135–139, 2012. doi:10.1126/science.1226518.

Pamela J Kennedy and Matthew L Shapiro. Motivational states activate distinct hippocampal representations to guide goal-directed behaviors. *Proceedings of the National Academy of Sciences*, 106(26):10805–10810, 2009. doi:10.1073/pnas.0903259106.

M. Keramati, A. Dezfouli, and P. Piray. Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput Biol*, 7(5):e1002055, 2011. doi:10.1371/journal.pcbi.1002055.

Raymond P Kesner and John C Churchwell. An analysis of rat prefrontal cortex in mediating executive function. *Neurobiology of learning and memory*, 96(3):417–431, 2011. doi:10.1016/j.nlm.2011.07.002.

S. Killcross and E. Coutureau. Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral cortex*, 13(4):400–408, 2003. doi:10.1093/cercor/13.4.400.

Ian Krajbich and Antonio Rangel. Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, 108(33):13852–13857, 2011. doi:10.1073/pnas.1101328108.

J. Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* Academic Press, 2014.

Donald Richard John Laming. *Information theory of choice-reaction times.* Academic Press, 1968.

Dylan M Layfield, Monica Patel, Henry Hallock, and Amy L Griffin. Inactivation of the nucleus reuniens/rhomboid causes a delay-dependent impairment of spatial working memory. *Neurobiology of learning and memory*, 125:163–167, 2015. doi:10.1016/j.nlm.2015.09.007.

S.W. Lee, S. Shimojo, and J.P. O'Doherty. Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3):687–699, 2014. doi:10.1016/j.neuron.2013.11.028.

Stefan Leutgeb, Jill K Leutgeb, Carol A Barnes, Edvard I Moser, Bruce L McNaughton, and May-Britt Moser. Independent codes for spatial and episodic memory in hippocampal neuronal ensembles. *Science*, 309(5734):619–623, 2005. doi:10.1126/science.1114037.

M.D. Lieberman. A social cognitive neuroscience approach. In *Social judgments: Implicit and explicit processes*, pages 44–67. Cambridge University Press, Cambridge UK, 2003.

Stephanie B Linley, Michelle M Gallo, and Robert P Vertes. Lesions of the ventral midline thalamus produce deficits in reversal learning and attention on an odor texture set shifting task. *Brain research*, 1649:110–122, 2016. doi:10.1016/j.brainres.2016.08.022.

Tiaotiao Liu, Wenwen Bai, Mi Xia, and Xin Tian. Directional hippocampal-prefrontal interactions during working memory. *Behavioural brain research*, 338:1–8, 2018. doi:10.1016/j.bbr.2017.10.003.

G. Loewenstein and T. O'Donoghue. Animal spirits: Affective and deliberative processes in economic behavior. *Working Paper; Available at SSRN 539843*, 2004.

Liya Ma, James M Hyman, Daniel Durstewitz, Anthony G Phillips, and Jeremy K Seamans. A quantitative analysis of context-dependent remapping of medial frontal cortex neurons and ensembles. *Journal of Neuroscience*, 36(31):8258–8272, 2016. doi:10.1523/JNEUROSCI.3176-15.2016.

Dennis M Maharjan, Yu Y Dai, Ethan H Glantz, and Shantanu P Jadhav. Disruption of dorsal hippocampal–prefrontal interactions using chemogenetic inactivation impairs spatial learning. *Neurobiology of learning and memory*, 155:351–360, 2018. doi:10.1016/j.nlm.2018.08.023.

David JN Maisson, Zachary M Gemzik, and Amy L Griffin. Optogenetic suppression of the nucleus reuniens selectively impairs encoding during spatial working memory. *Neurobiology of Learning and Memory*, 155:78–85, 2018. doi:10.1016/j.nlm.2018.06.010.

Hugo Malagon-Vina, Stephane Ciocchi, Johannes Passecker, Georg Dorffner, and Thomas Klausberger. Fluid network dynamics in the prefrontal cortex during multiple strategy switching. *Nature communications*, 9(1):309, 2018. doi:10.1038/s41467-017-02764-x.

Emily A Mankin, Fraser T Sparks, Begum Slayyeh, Robert J Sutherland, Stefan Leutgeb, and Jill K Leutgeb. Neuronal code for extended time in the hippocampus. *Proceedings of the National Academy of Sciences*, 109(47):19462–19467, 2012. doi:10.1073/pnas.1214107109.

Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503 (7474):78, 2013. doi:10.1038/nature12742.

Jean-Philippe Marquis, Simon Killcross, and Josephine E Haddon. Inactivation of the prelimbic, but not infralimbic, prefrontal cortex impairs the contextual control of response conflict in rats. *European Journal of Neuroscience*, 25(2):559–566, 2007. doi:10.1111/j.1460-9568.2006.05295.x.

Kenji Matsumoto and Keiji Tanaka. The role of the medial prefrontal cortex in achieving goals. *Current opinion in neurobiology*, 14(2):178–185, 2004. doi:10.1016/j.conb.2004.03.005.

Kenji Matsumoto, Wataru Suzuki, and Keiji Tanaka. Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science*, 301(5630):229–232, 2003. doi:10.1126/science.1084204.

James Timothy McKenna and Robert P Vertes. Afferent projections to nucleus reuniens of the thalamus. *Journal of comparative neurology*, 480(2):115–142, 2004. doi:10.1002/cne.20342.

Hao Mei, Nikos K Logothetis, and Oxana Eschenko. The activity of thalamic nucleus reuniens is critical for memory retrieval, but not essential for the early phase of “off-line” consolidation. *Learning & Memory*, 25(3):129–137, 2018. doi:10.1101/lm.047134.117.

Ger-Jan Mensink and Jeroen G Raaijmakers. A model for interference and forgetting. *Psychological Review*, 95(4):434, 1988. doi:10.1037/0033-295X.95.4.434.

Earl K Miller and Jonathan D Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001. doi:10.1146/annurev.neuro.24.1.167.

K.J. Miller, J.C. Erlich, C.D. Kopec, M.M. Botvinick, and C.D. Brody. A multi-step decision task to distinguish model-based from model-free reinforcement learning in rats. Presentation at the Society for Neuroscience Annual Meeting, San Diego, CA. Program No. 855.13., 2013.

K.J. Miller, J. Erlich, C. Kopec, M. Botvinick, and C. Brody. A multi-step decision task elicits planning behavior in rats. Presentation at the Computational and Systems Neuroscience Conference (Cosyne), Salt Lake City, UT, 2014.

K.J. Miller, M.M. Botvinick, and C.D. Brody. Dorsal hippocampus contributes to model-based planning. *Nature Neuroscience*, 20(9):1269–1276, 2017. doi:10.1038/nn.4613.

P Mitra, H. Bokil, H Maniar, C Loader, S Mehta, D Hill, S Mitra, P Andrews, R Baptista, S Gopinath, H Nalatore, and S Kaur. Chronux. Version 2.12 v03, 2018. URL `http://chronux.org`.

Partha Mitra. *Observed brain dynamics*. Oxford University Press, 2007.

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

K. Muenzinger and E. Gentry. Tone discrimination in white rats. *Journal of Comparative Psychology*, 12(2):195, 1931. doi:10.1037/h0072238.

Nandakumar S Narayanan and Mark Laubach. Neuronal correlates of post-error slowing in the rat dorsomedial prefrontal cortex. *Journal of neurophysiology*, 100 (1):520–525, 2008. doi:10.1152/jn.00035.2008.

Rapeechai Navawongse and Howard Eichenbaum. Distinct pathways for rule-based retrieval and spatial mapping of memory representations in hippocampal neurons. *Journal of Neuroscience*, 33(3):1002–1013, 2013. doi:10.1523/JNEUROSCI.3891-12.2013.

Y. Niv, D. Joel, and P. Dayan. A normative perspective on motivation. *Trends in cognitive sciences*, 10(8):375–381, 2006. doi:10.1016/j.tics.2006.06.010.

Elisabeth Obst, Daniel J Schad, Quentin JM Huys, Miriam Sebold, Stephan Nebe, Christian Sommer, Michael N Smolka, and Ulrich S Zimmermann. Drunk decisions: Alcohol shifts choice from habitual towards goal-directed control in adolescent intermediate-risk drinkers. *Journal of psychopharmacology*, 32(8):855–866, 2018. doi:10.1177/0269881118772454.

J. O'Keefe and L. Nadel. *The hippocampus as a cognitive map.* Oxford University Press, USA, 1978a.

J O'Keefe and L Nadel. *The Hippocampus as a Cognitive Map.* Oxford University Press, 1978b.

John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 1971. doi:10.1016/0006-8993(71)90358-1.

Pia-Kelsey O'Neill, Joshua A Gordon, and Torfi Sigurdsson. Theta oscillations in the medial prefrontal cortex are modulated by spatial working memory and synchronize with the hippocampus through its ventral subregion. *Journal of Neuroscience*, 33 (35):14211–14224, 2013. doi:10.1523/JNEUROSCI.2378-13.2013.

A.R. Otto, S.J. Gershman, A.B. Markman, and N.D. Daw. The curse of planning dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological science*, 24(5):751–761, 2013a. doi:10.1177/0956797612463080.

A.R. Otto, C.M. Raio, A. Chiang, E.A. Phelps, and N.D. Daw. Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, 110(52):20941–20946, 2013b. doi:10.1073/pnas.1312011110.

A.R. Otto, A. Skatova, S. Madlon-Kay, and N.D. Daw. Cognitive control predicts use of model-based reinforcement learning. *Journal of Cognitive Neuroscience*, 27 (2):319–333, 2015. doi:10.1162/jocn_a_00709.

Marius Pachitariu, Nicholas A Steinmetz, Shabnam N Kadir, Matteo Carandini, and Kenneth D Harris. Fast and accurate spike sorting of high-channel count probes with kilosort. In *Advances in Neural Information Processing Systems*, pages 4448–4456, 2016.

M.G. Packard and J.L. McGaugh. Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of learning and memory*, 65(1):65–72, 1996. doi:10.1006/nlme.1996.0007.

Camillo Padoa-Schioppa. Neurobiology of economic choice: a good-based model. *Annual review of neuroscience*, 34:333–359, 2011. doi:10.1146/annurev-neuro-061010-113648.

Andrew E Papale, Jeffrey J Stott, Nathaniel J Powell, Paul S Regier, and A David Redish. Interactions between deliberation and delay-discounting in rats. *Cognitive, Affective, & Behavioral Neuroscience*, 12(3):513–526, 2012. doi:10.3758/s13415-012-0097-7.

George Paxinos and Charles Watson. *The rat brain in stereotaxic coordinates: hard cover edition*. Elsevier, 2006.

Ryan Place, Anja Farovik, Marco Brockmann, and Howard Eichenbaum. Bidirectional prefrontal-hippocampal interactions support context-guided memory. *Nature neuroscience*, 19(8):992, 2016. doi:10.1038/nn.4327.

Nathaniel J Powell and A David Redish. Complex neural codes in rat prelimbic cortex are stable across days on a spatial decision task. *Frontiers in behavioral neuroscience*, 8:120, 2014. doi:10.3389/fnbeh.2014.00120.

Nathaniel James Powell and A David Redish. Representational changes of latent strategies in rat medial prefrontal cortex precede changes in behaviour. *Nature communications*, 7:12830, 2016. doi:10.1038/ncomms12830.

Alison R Preston and Howard Eichenbaum. Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, 23(17):R764–R773, 2013. doi:10.1016/j.cub.2013.05.041.

C. Radenbach, A.M. Reiter, V. Engert, Z. Sjoerds, A. Villringer, H.J. Heinze, L. Deserno, and F. Schlagenhauf. The interaction of acute and chronic stress impairs model-based behavioral control. *Psychoneuroendocrinology*, 53:268–280, 2015. doi:10.1016/j.psyneuen.2014.12.017.

Michael E Ragozzino and Raymond P Kesner. The effects of muscarinic cholinergic receptor blockade in the rat anterior cingulate and prelimbic/infralimbic cortices on spatial working memory. *Neurobiology of learning and memory*, 69(3):241–257, 1998. doi:10.1006/nlme.1998.3823.

Michael E Ragozzino, Jenna Kim, Derrick Hassert, Nancy Minniti, and Charlene Kiang. The contribution of the rat prelimbic-infralimbic areas to different forms of task switching. *Behavioral neuroscience*, 117(5):1054, 2003. doi:10.1037/0735-7044.117.5.1054.

Antonio Rangel, Colin Camerer, and P Read Montague. A framework for studying the neurobiology of value-based decision making. *Nature reviews neuroscience*, 9 (7):545, 2008. doi:10.1038/nrn2357.

Roger Ratcliff. A theory of memory retrieval. *Psychological review*, 85(2):59, 1978. doi:10.1037/0033-295X.85.2.59.

Roger Ratcliff and Gail McKoon. The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, 20(4):873–922, 2008. doi:10.1162/neco.2008.12-06-420.

A David Redish. Mclust. Version 4.4.07, 2017. URL `http://redishlab.neuroscience.umn.edu/MClust/MClust.html`.

A.D. Redish. *Beyond the cognitive map: from place cells to episodic memory*. MIT Press, 1999.

A.D. Redish. *The mind within the brain: How we make decisions and how those decisions go wrong.* Oxford University Press, 2013.

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

A.D. Redish, S. Jensen, A. Johnson, and Z. Kurth-Nelson. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological review*, 114(3):784, 2007. doi:10.1037/0033-295X.114.3.784.

Paul S Regier, Seiichiro Amemiya, and A David Redish. Hippocampus and subregions of the dorsal striatum respond differently to a behavioral strategy change on a spatial navigation task. *Journal of neurophysiology*, 114(3):1399–1416, 2015a. doi:10.1152/jn.00189.2015.

P.S. Regier, S. Amemiya, and A.D. Redish. Decision making: Neural mechanisms: Hippocampus and subregions of the dorsal striatum respond differently to a behavioral strategy change on a spatial navigation task. *Journal of Neurophysiology*, 114 (3):1399, 2015b. doi:10.1152/jn.00189.2015.

E.L. Rich and J.D. Wallis. Decoding subjective decisions from orbitofrontal cortex. *Nature neuroscience*, 19(7):973, 2016. doi:10.1038/nn.4320.

Erin L Rich and Matthew Shapiro. Rat prefrontal cortical neurons selectively code strategy switches. *Journal of Neuroscience*, 29(22):7208–7219, 2009. doi:10.1523/JNEUROSCI.6068-08.2009.

Cyrille Rossant, Shabnam Kadir, Dan Goodman, Max Hunter, and Kenneth Harris. Phy. Version 1.0.9, 2016. URL `http://github.com/cortex-lab/phy`.

Alex Roxin and Stefano Fusi. Efficient partitioning of memory systems and its importance for memory consolidation. *PLoS computational biology*, 9(7):e1003146, 2013. doi:10.1371/journal.pcbi.1003146.

Alon Rubin, Nitzan Geva, Liron Sheintuch, and Yaniv Ziv. Hippocampal ensemble dynamics timestamp events in long-term memory. *Elife*, 4:e12247, 2015. doi:10.7554/eLife.12247.

Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*. University of Cambridge, Department of Engineering, 1994.

D.J. Schad, E. Jünger, M. Sebold, M. Garbusow, N. Bernhardt, A. Javadi, U.S. Zimmermann, M.N. Smolka, A. Heinz, M.A. Rapp, and Q.J.M. Huys. Processing speed enhances model-based over model-free reinforcement learning in the presence of high working memory functioning. *Frontiers in psychology*, 5, 2014. doi:10.3389/fpsyg.2014.01450.

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

Brandy Schmidt, Anneke A Duin, and A David Redish. Disrupting the medial prefrontal cortex alters hippocampal sequences during deliberative decision-making. *Journal of neurophysiology*, 2019. doi:10.1152/jn.00793.2018.

N. Schmitzer-Torbert and A.D. Redish. Development of path stereotypy in a single day in rats on a multiple-t maze. *Archives italiennes de biologie*, 140(4):295–301, 2002. doi:10.4449/aib.v140i4.488.

N Schmitzer-Torbert, J Jackson, D Henze, K Harris, and AD Redish. Quantitative measures of cluster quality for use in extracellular recordings. *Neuroscience*, 131 (1):1–11, 2005. doi:10.1016/j.neuroscience.2004.09.066.

Neil Schmitzer-Torbert and A David Redish. Neuronal activity in the rodent dorsal striatum in sequential navigation: separation of spatial and reward responses on the multiple t task. *Journal of neurophysiology*, 91(5):2259–2272, 2004. doi:10.1152/jn.00687.2003.

W. Schultz, P. Dayan, and P.R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997. doi:10.1126/science.275.5306.1593.

W.B. Scoville and B. Milner. Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry*, 20(1):11, 1957. doi:10.1136/jnnp.20.1.11.

Jeremy K Seamans, Stanley B Floresco, and Anthony G Phillips. Functional differences between the prelimbic and anterior cingulate regions of the rat prefrontal cortex. *Behavioral neuroscience*, 109(6):1063, 1995. doi:10.1037/0735-7044.109.6.1063.

M. Sebold, L. Deserno, S. Nebe, D.J. Schad, M. Garbusow, C. Hägele, J. Keller, E. Jünger, N. Kathmann, M. Smolka, M.A. Rapp, F. Schlagenhauf, A. Heinz, and Q.J.M. Huys. Model-based and model-free decisions in alcohol dependence. *Neuropsychobiology*, 70(2):122–131, 2014. doi:10.1159/000362840.

M.E. Sharp, K. Foerde, N.D. Daw, and D. Shohamy. Dopamine selectively remediates 'model-based' reward learning: a computational approach. *Brain*, 139(2):355–364, 2015. doi:10.1093/brain/awv347.

Patricia E Sharp, Hugh T Blair, David Etkin, and Douglas B Tzanetos. Influences of vestibular and visual motion information on the spatial firing patterns of hippocampal place cells. *Journal of Neuroscience*, 15(1):173–189, 1995. doi:10.1523/JNEUROSCI.15-01-00173.1995.

PE Sharp, JL Kubie, and RU Muller. Firing properties of hippocampal neurons in a visually symmetrical environment: contributions of multiple sensory cues and mnemonic processes. *Journal of Neuroscience*, 10(9):3093–3105, 1990. doi:10.1523/JNEUROSCI.10-09-03093.1990.

Justin D Shin and Shantanu P Jadhav. Multiple modes of hippocampal–prefrontal interactions in memory-guided behavior. *Current opinion in neurobiology*, 40:161–169, 2016. doi:10.1016/j.conb.2016.07.015.

Athanassios G Siapas, Evgueniy V Lubenov, and Matthew A Wilson. Prefrontal phase locking to hippocampal theta oscillations. *Neuron*, 46(1):141–151, 2005. doi:10.1016/j.neuron.2005.02.028.

D.A. Simon and N.D. Daw. Neural correlates of forward planning in a spatial decision task in humans. *The Journal of Neuroscience*, 31(14):5526–5539, 2011. doi:10.1523/JNEUROSCI.4647-10.2011.

Anton Sirota, Sean Montgomery, Shigeyoshi Fujisawa, Yoshikazu Isomura, Michael Zugaro, and György Buzsáki. Entrainment of neocortical neurons and gamma oscillations by the hippocampal theta rhythm. *Neuron*, 60(4):683–697, 2008. doi:10.1016/j.neuron.2008.09.014.

A. Skatova, P.A. Chan, and N.D. Daw. Extraversion differentiates between model-based and model-free strategies in a reinforcement learning task. *Frontiers in Human Neuroscience*, 7(525), 2013. doi:10.3389/fnhum.2013.00525.

S.A. Sloman. The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1):3, 1996. doi:10.1037/0033-2909.119.1.3.

David M Smith and Sheri JY Mizumori. Learning-related development of context-specific neuronal responses to places and events: the hippocampal role in context processing. *Journal of Neuroscience*, 26(12):3154–3163, 2006. doi:10.1523/JNEUROSCI.3234-05.2006.

K.S. Smith and A.M. Graybiel. A dual operator view of habitual behavior reflecting cortical and striatal dynamics. *Neuron*, 79(2):361–374, 2013. doi:10.1016/j.neuron.2013.05.038.

Timothy Spellman, Mattia Rigotti, Susanne E Ahmari, Stefano Fusi, Joseph A Gogos, and Joshua A Gordon. Hippocampal–prefrontal input supports spatial encoding in working memory. *Nature*, 522(7556):309, 2015. doi:10.1038/nature14445.

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002. doi:10.1111/1467-9868.00353.

Jennifer R St. Onge and Stan B Floresco. Prefrontal cortical contribution to risk-based decision making. *Cerebral cortex*, 20(8):1816–1828, 2009. doi:10.1093/cercor/bhp250.

Stan Development Team. Pystan: the python interface to stan, 2017. URL `http://mc-stan.org`. Version 2.16.0.0.

Adam P Steiner and A David Redish. The road not taken: neural correlates of decision making in orbitofrontal cortex. *Frontiers in neuroscience*, 6:131, 2012. doi:10.3389/fnins.2012.00131.

Mervyn Stone. Models for choice-reaction time. *Psychometrika*, 25(3):251–260, 1960. doi:10.1007/BF02289729.

Jeffrey J Stott and A David Redish. 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, 2014. doi:10.1098/rstb.2013.0472.

Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4):160–163, 1991. doi:10.1145/122344.122377.

R.S. Sutton and A.G. Barto. *Reinforcement learning: An introduction.* MIT press Cambridge, 1998.

LW Swanson. A direct projection from ammon's horn to prefrontal cortex in the rat. *Brain research*, 217(1):150–154, 1981. doi:10.1016/0006-8993(81)90192-X.

L.W. Swanson. Cerebral hemisphere regulation of motivated behavior. *Brain research*, 886(1):113–164, 2000. doi:10.1016/S0006-8993(00)02905-X.

Andreas S Tolias, Alexander S Ecker, Athanassios G Siapas, Andreas Hoenselaar, Georgios A Keliris, and Nikos K Logothetis. Recording chronically from the same neurons in awake, behaving primates. *Journal of neurophysiology*, 98(6):3780–3790, 2007. doi:10.1152/jn.00260.2007.

E.C. Tolman. Prediction of vicarious trial and error by means of the schematic sowbug. *Psychological Review*, 46(4):318, 1939. doi:10.1037/h0057054.

Sophie Tronel and Susan J Sara. Blockade of nmda receptors in prelimbic cortex induces an enduring amnesia for odor–reward associative learning. *Journal of Neuroscience*, 23(13):5472–5476, 2003. doi:10.1523/JNEUROSCI.23-13-05472.2003.

Kimberly R Urban, Dylan M Layfield, and Amy L Griffin. Transient inactivation of the medial prefrontal cortex impairs performance on a working memory-dependent conditional discrimination task. *Behavioral neuroscience*, 128(6):639, 2014. doi:10.1037/bne0000020.

M. van der Meer, Z. Kurth-Nelson, and A.D. Redish. Information processing in decision-making systems. *The Neuroscientist*, 18(4):342–359, 2012. doi:10.1177/1073858411435128.

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

Robert P Vertes. Analysis of projections from the medial prefrontal cortex to the thalamus in the rat, with emphasis on nucleus reuniens. *Journal of Comparative Neurology*, 442(2):163–187, 2002. doi:10.1002/cne.10083.

Robert P Vertes. Differential projections of the infralimbic and prelimbic cortex in the rat. *Synapse*, 51(1):32–58, 2004. doi:10.1002/syn.10279.

Robert P Vertes, Walter B Hoover, Angela Cristina Do Valle, Alexandra Sherman, and JJ Rodriguez. Efferent projections of reuniens and rhomboid nuclei of the thalamus in the rat. *Journal of comparative neurology*, 499(5):768–796, 2006. doi:10.1002/cne.21135.

Robert P Vertes, Walter B Hoover, Klara Szigeti-Buck, and Csaba Leranth. Nucleus reuniens of the midline thalamus: link between the medial prefrontal cortex and the hippocampus. *Brain research bulletin*, 71(6):601–609, 2007. doi:10.1016/j.brainresbull.2006.12.002.

Ronald WH Verwer, Ron J Meijer, Hannie FM Van Uum, and Menno P Witter. Collateral projections from the rat hippocampal formation to the lateral and medial prefrontal cortex. *Hippocampus*, 7(4):397–402, 1997. doi:10.1002/(SICI)1098-1063(1997)7:4<397::AID-HIPO5>3.0.CO;2-G.

Tatiana D Viena, Stephanie B Linley, and Robert P Vertes. Inactivation of nucleus reuniens impairs spatial working memory and behavioral flexibility in the rat. *Hippocampus*, 28(4):297–311, 2018. doi:10.1002/hipo.22831.

V. Voon, K. Derbyshire, C. Rück, M.A. Irvine, Y. Worbe, J. Enander, L.R.N. Schreiber, C. Gillan, N.A. Fineberg, B.J. Sahakian, T.W. Robbins, N.A. Harrison, J. Wood, N.D. Daw, P. Dayan, J.E. Grant, and E.T. Bullmore. Disorders of compulsivity: a common bias towards learning habits. *Molecular psychiatry*, 20(3): 345–352, 2015. doi:10.1038/mp.2014.44.

J.D. Wallis. Decoding cognitive processes from neural ensembles. *Trends in cognitive sciences*, 22(12):1091–1102, 2018. doi:10.1016/j.tics.2018.09.002.

Jonathan D Wallis, Kathleen C Anderson, and Earl K Miller. Single neurons in prefrontal cortex encode abstract rules. *Nature*, 411(6840):953, 2001. doi:10.1038/35082081.

Gong-Wu Wang and Jing-Xia Cai. Disconnection of the hippocampal–prefrontal cortical circuits impairs spatial working memory performance in rats. *Behavioural brain research*, 175(2):329–336, 2006. doi:10.1016/j.bbr.2006.09.002.

Gong-Wu Wang and Jing-Xia Cai. Reversible disconnection of the hippocampal-prelimbic cortical circuit impairs spatial learning but not passive avoidance learning in rats. *Neurobiology of learning and memory*, 90(2):365–373, 2008. doi:10.1016/j.nlm.2008.05.009.

Jane X Wang, Neal J Cohen, and Joel L Voss. Covert rapid action-memory simulation (crams): A hypothesis of hippocampal–prefrontal interactions for adaptive behavior. *Neurobiology of learning and memory*, 117:22–33, 2015. doi:10.1016/j.nlm.2014.04.003.

C.J.C.H. Watkins. *Learning from delayed rewards.* PhD thesis, King's College, Cambridge, Cambridge, 1989.

A.M. Wikenheiser and A.D. Redish. Hippocampal theta sequences reflect current goals. *Nature neuroscience*, 18(2):289–294, 2015. doi:10.1038/nn.3909.

A.M. Wikenheiser and G. Schoenbaum. Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. *Nature Reviews Neuroscience*, 17(8):513–523, 2016. doi:10.1038/nrn.2016.56.

Emma R Wood, Paul A Dudchenko, R Jonathan Robitsek, and Howard Eichenbaum. Hippocampal neurons encode information about different types of memory episodes occurring in the same location. *Neuron*, 27(3):623–633, 2000. doi:10.1016/S0896-6273(00)00071-4.

K. Wunderlich, P. Smittenaar, and R.J. Dolan. Dopamine enhances model-based over model-free choice behavior. *Neuron*, 75(3):418–424, 2012. doi:10.1016/j.neuron.2012.03.042.

Wei Xu and Thomas C Südhof. A neural circuit for memory specificity and generalization. *Science*, 339(6125):1290–1295, 2013. doi:10.1126/science.1229534.

H.H. Yin and B.J. Knowlton. Contributions of striatal subregions to place and response learning. *Learning & Memory*, 11(4):459–463, 2004. doi:10.1101/lm.81004.

Taejib Yoon, Jeffrey Okada, Min W Jung, and Jeansok J Kim. Prefrontal cortex and hippocampus subserve different components of working memory in rats. *Learning & memory*, 15(3):97–105, 2008. doi:10.1101/lm.850808.

James J Young and Matthew L Shapiro. Double dissociation and hierarchical organization of strategy switches and reversals in the rat pfc. *Behavioral neuroscience*, 123(5):1028, 2009. doi:10.1037/a0016822.

Kechen Zhang, Iris Ginzburg, Bruce L McNaughton, and Terrence J Sejnowski. Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. *Journal of neurophysiology*, 79(2):1017–1044, 1998. doi:10.1152/jn.1998.79.2.1017.

J. Zhou, M. Montesinos-Cartagena, A.M. Wikenheiser, M.P.H. Gardner, Y. Niv, and G. Schoenbaum. Complementary task structure representations in hippocampus and orbitofrontal cortex during an odor sequence task. *Current Biology*, 29(20): 3402–3409, 2019. doi:10.1016/j.cub.2019.08.040.

Mark C Zielinski, Justin D Shin, and Shantanu P Jadhav. Coherent coding of spatial position mediated by theta oscillations in the hippocampus and prefrontal cortex. *Journal of Neuroscience*, 39(23):4550–4565, 2019. doi:10.1523/JNEUROSCI.0106-19.2019.

Eric A Zilli and Michael E Hasselmo. Modeling the role of working memory and episodic memory in behavioral tasks. *Hippocampus*, 18(2):193–209, 2008. doi:10.1002/hipo.20382.

Eric C Zimmerman and Anthony A Grace. Prefrontal cortex modulates firing pattern in the nucleus reuniens of the midline thalamus via distinct corticothalamic pathways. *European Journal of Neuroscience*, 48(10):3255–3272, 2018. doi:10.1111/ejn.14111.

Yaniv Ziv, Laurie D Burns, Eric D Cocker, Elizabeth O Hamel, Kunal K Ghosh, Lacey J Kitch, Abbas El Gamal, and Mark J Schnitzer. Long-term dynamics of ca1 hippocampal place codes. *Nature neuroscience*, 16(3):264, 2013. doi:10.1038/nn.3329.