Functions of the hippocampal memory system in instrumental control

by

Aaron M. Bornstein

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Psychology

New York University

January 2014

Dr. Nathaniel D. Daw

“Strictly speaking, we do not make decisions, decisions make us.”

– José Saramago, *All the Names*

“There is a great deal of chance in all this...”

– Marcel Proust, *Swann’s Way*

ABSTRACT

That our decisions are guided by internal representations of our environments is one of the founding observations of cognitive psychology — the celebrated 'cognitive map' of Tolman. But where in the brain is this map represented, and how is it used to make decisions? A significant body of neuroscientific evidence points to the hippocampal memory system as critical for the internal representation of environmental structure. However, this work has been mainly indirect in nature — demonstrating the influence of hippocampal representations, but not their direct use in decisions. In parallel, computational reinforcement learning has addressed the problem of learning environmental structure to efficiently guide decisions, within constraints of information storage and processing capability. This work is known as 'model-based' reinforcement learning, in contrast to the 'model-free' variety used to describe action-value learning in the mesostriatal dopamine system. A key heuristic emerging from this literature is the notion that 'sampling' from past experiences can be used to estimate values, and guide actions, at the time of decision. Here, we join these two literatures — the neuroscientific study of the hippocampal memory system in decisions, and the computational study of reinforcement learning under resource constraints — to explore the neural realization of model-based decisions. Specifically, we show the hippocampal memory system is implicated in representing sequential contingencies in a serial learning task. Next, this representation is shown to directly impact decisions in an embedded task where choices crucially rely on the learned sequential contingencies. Our results suggest that these choices are made after estimating action values by sampling from episodes of previ-

ous experience. Finally, we investigate this sampling mechanism by using a modified choice task to causally manipulate the availability of episodes for making choices. We show that 'priming' past episodes, using task-irrelevant memory cues, can in fact bias subsequent choices to be guided by the reward values associated with those primed experiences. This finding suggests an alternative mechanism to the standard account of strict recency-weighted value estimation. Taken together, these studies outline a role for the hippocampus in instrumental control, and open numerous questions for further research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 has been published elsewhere (European Journal of Neuroscience, 2012). Per the copyright agreement with European Journal of Neuroscience, authors may use their own published material in other publications.

INTRODUCTION

The decisions we make rely on our knowledge of the world, and, more specifically, our understanding of the contexts in which we act. The layout of space, the temporal order of events, the implied presence of one object when viewing another — all of these are components of our context, and factor into the ways we interact with the world. Models that leverage one particular type of context, temporal recency, have proven immensely valuable in understanding the neural substrates of decisions based on trial-and-error learning about the reinforcement value of our actions. But we are guided by more than just incremental learning of action values: humans and animals can assemble from our experiences a complex understanding of the contingencies in the world around us, and use those contingencies to plan and act. That our decisions could critically incorporate such a "cognitive map" of contingencies in the world was among the founding observations of cognitive psychology (Tolman, 1948). Tolman's elegant demonstrations set forth the idea that learned information can be recombined — generalized across contexts — and thus *flexibly* expressed to achieve novel goals. The notion that action can be guided by rich internal representations forms the basis of many ideas in research on both memory and decisions.

Since that time, neuroscientific investigations building from this proposal have yielded enormous insight into the mechanisms of action selection, the processes that underlie decisions for rewards, and the representations of environmental structure and episodic experience. Recently, normative approaches gleaned from computer science and economics have proved useful tools for weaving together neuroscientific data into a bur-

geoning theory of how values are learned, represented, and transformed into action. These approaches have primarily been useful in describing decisions resulting from a history of direct reinforcement — known as habitual or 'model-free' learning. However, large gaps remain in our knowledge of the mechanisms underlying a broad range of human and animal decisions for rewards. We have in particular a poor understanding of precisely how — at a computational, psychological, and neural level — we make those decisions that exploit a fully featured model of contingencies. These types of decisions are variously termed 'goal-directed' or 'model-based', to reflect the fact that they rely on representations more expansive than reinforced habits.

A leading candidate for the neural substrate of the cognitive map is the medial temporal lobe memory system, centered on the hippocampus. The features of representations required for goal-directed decisions (Balleine and Dickinson, 1998) are precisely those long associated with the memories formed by this system (Balleine et al., 2008). Namely, they can be *flexibly expressed* — that is, associated with other memories, in whole or in part — and so, they enable *outcome-sensitive* and *context-aware* behaviors — actions can be chosen based on specific features of the outcome, not just the expected reward values, and associations can be selected based on their appropriateness to the current context (e.g., satiety, latent causes). Model-based control — the computational reinforcement learning parallel of goal-directed control (Balleine et al., 2008; Bornstein and Daw, 2011) — may be accomplished by either evaluating structured models of the associations in the environment (e.g., decision trees Daw et al. 2005), or by drawing on episodes — called 'samples' — of past experience (Lengyel and Dayan, 2008). (This latter approach, treating episodes as units of evidence about the structure

of the world, will be a primary focus of the experiments presented here.) Normal acquisition of both of these types of memories is thought to be critically dependent on the hippocampal memory system (Squire, 1992; McClelland et al., 1995). Therefore, mechanisms for the representation and retrieval of hippocampal system memories may be crucial to model-based control.

In this dissertation I discuss how the hippocampus and related structures may support model-based control of behavior. I provide new experimental evidence substantiating the role of the hippocampal memory system in normal instrumental action. Additionally, I argue that situating the hippocampus within the mainstream analysis of reward-guided decision making has the potential to provide a unifying explanation of many heretofore puzzling features of choice.

In this section I review computational and neuroscientific studies of model-based reinforcement learning. I also review empirical and theoretical work outlining functions of the hippocampus in learning, memory, and decision-making. I place particular emphasis on studies of hippocampal activity in reward-guided decisions with shifting or uncertain contingencies, the precise situations in which model-based control is most desirable.

Model-based reinforcement learning

In this section I introduce model-based reinforcement learning. Specifically, I outline the extant computational approaches to model-based reinforcement learning, and identify parallels with previously described neurobiological mechanisms for learning and choice. Beginning with concepts from the study of instrumental control in humans and animals, I describe the computational distinction between model-free and model-based

reinforcement learning as critically hinging on how learned information is represented, and what types of uses those representations enable. I then review neuroscientific studies in humans and animals that examine the biological underpinnings of these representations.

Background

The control of behavior is critically dependent on what we learn about the consequences of our actions and the structure of the world, and how those *contingencies* are represented. The study of instrumental learning classifies behavior among several subtypes, distinguished by the representations used to guide actions.

In the analysis of how learning supports decisions, researchers model actors as choosing from among a set of actions, each of which is expected to lead to some amount of reinforcement (or punishment). Actors are expected to choose, with some high probability, the action that they expect to be most rewarding. (The proportion of times that the most highly rewarded action is selected is often dependent on the difference between the reward expected from that action and the reward expected from other actions — this 'soft max' approach enables 'probability matching' and rudimentary exploratory behaviors.) This key piece of information, the *decision variable* (the value of actions), can be learned by trial and error, or estimated at the time of choice, according to one of a variety of approaches. Which approach is used depends critically on the type of *representation* employed.

Often, this research evaluates tasks that have additional structure, besides the rewards associated with actions. For example, sequential tasks might have dependencies that extend over multiple trials: I might choose to enter a room of a maze and not be

able to turn back, or I might be playing a competitive game such as chess, in which my moves change the state of the world. These dependencies that link 'states' of the environment — the collection of which are known as a 'world model' — may also be learned by the actor. Again, the types of decisions made in these tasks depend on the types of representations learned, and how they are used to select actions.

'Habitual' actions are based on solely stimulus-response associations, while 'goal-directed' actions are supported by a representation of the particular goal (such as a type of food) expected for an action — crucially, this goal representation is flexibly linked to the actions that might bring it about (Balleine, 2005). This distinction is typically probed using manipulations that alter the action-outcome contingency or outcome value, such as reward devaluation, and then evaluating the subsequent effect on responding. Early in learning, animals may shift behavior to adapt to these kinds of changes in the goal, or to changes in the environment that alter the sequence of steps they need to take to achieve the goal. After extensive training, behavior can become insensitive to these manipulations, suggesting that the animal has transitioned to an isolated reliance on stimulus-response representations — that is, *habits*.

Different types of learning are proposed to encode these different representations, and entail distinct forms of value estimation. In the case of stimulus-response learning, the value estimation process occurs at the time of encoding the representation. That is, when the experienced reward is received, it is used to potentiate the preceding action. In the more broad class of representations used to support goal-directed control, values cannot, by definition, be estimated at the time of learning. Instead, associations are learned between many features of the environment, only to be later used to dynami-

cally construct a value estimate by identifying which of those learned contingencies are relevant to the receipt of reward. The mechanisms by which this occurs are not fully specified, and it is likely that there are many.

Extensive theoretical and experimental work has described neural mechanisms by which the values of actions are updated via direct experience. This work has primarily focused on the mesostriatal dopamine system. Particularly fruitful has been the observation that activity of dopamine neurons carries a signature of a family of computational algorithms known collectively as temporal-difference learning (TD Watkins, 1989; Barto and Sutton, 1998). Specifically, phasic dopamine responses qualitatively (Montague et al., 1996; Schultz et al., 1997) and quantitatively (Houk et al., 1995; Bayer and Glimcher, 2005) match the reward prediction error (RPE) — the difference between the amount of reward received and that expected — computed by these algorithms. In the well-studied 'actor/critic' architecture (Barto and Sutton, 1998; O'Doherty et al., 2004) this signal is used to reinforce action tendencies that are thought to be critically dependent on the dorsolateral striatum (Yin et al., 2005b).

Giving tangibility to this idea, and situating it within the broader context of coexisting goal-directed and habitual control systems, are a number of studies building on a key discovery made by interrupting the function of subregions of rodent dorsal striatum (Yin et al., 2004, 2005b, 2008; Balleine et al., 2009). In this seminal series of experiments, Yin and colleagues observed that lesions of dorsolateral striatum and pharmacological disruptions of dorsomedial striatum (DLS and DMS) specifically disabled habitual and goal-directed responding, respectively (Yin et al., 2004, 2005b,a).

A similar distinction has emerged in the literature on computational theories of rein-

forcement learning (Doya, 1999; Daw et al., 2005; Niv et al., 2006; Redish et al., 2008; Rangel et al., 2008; Balleine et al., 2008; Frank et al., 2009). Specifically, it is well recognized that while 'model-free' TD architectures like actor/critic successfully describe stimulus-response habit learning (Suri and Schultz, 2001), they fail to capture characteristics of goal-directed choice like devaluation sensitivity and latent learning (Tolman, 1948). Therefore, model-based RL, and its ability to evaluate potential actions at the time of choice, is often associated with goal-directed choice, and thus its corresponding neural substrates in DMS (Bornstein and Daw, 2011).

While this mapping has much promise for fostering computational modeling of goal-directed behavior, in the same way that such models have yielded insights about habit learning, key questions remain. For one, there are several ways to evaluate cognitive models — such as probabilistic inference (Solway and Botvinick, 2012), or simulation of a forward model if one is available (Johnson et al., 2007; Addis et al., 2007). A common feature of both of these types of evaluation is that they can construct new action plans, or simply re-evaluate old ones, by drawing on information besides the history of reward. This is in fact the critical signature of model-based RL, one it shares with goal-directed learning. Thus, however it is implemented, the use of contingencies learned in the absence of reward is a necessary distinguishing trait of model-based control.

A number of different approaches have been proposed for how these contingencies might be learned and used. Recently, work has attempted to identify a neural prediction error signature, analogous to the RPE, of updating the representations used in model-based learning — a 'state prediction error' (SPE) (Hampton et al., 2006, 2008; Gläscher et al., 2010). Perhaps relatedly, a reported hippocampal 'novelty signal' (Lisman and

Grace, 2005; Kumaran and Maguire, 2006; Duncan et al., 2011) might indicate acquisition of the sequential representations of the kind useful to model-based control. It remains difficult to distinguish between learning, or novelty detection, *per se*, and learning which is directed towards acquiring associations for later goal-directed decisions. Partly due to the ill-formed nature of the endeavor of identifying general features of models of environmental structure, no consensus has yet been achieved on the neural mechanisms of model learning.

An inference approach which relies on sampling discrete episodes of past experience might sidestep this problem of proliferating prediction errors. Instead of updating a value estimate, or contingency representation, these algorithms populate a cache of episodes as they are experienced, with no explicit reference to expectations (and thus, no prediction error). However a purely sample-driven approach entails a separate question, which is how the reward prediction error is computed *at all* — and why it appears to be influenced by model-based expectations (Gläscher et al., 2010; Daw et al., 2011; Diuk et al., 2013). It may be that certain tasks preferentially engage the strategy of sampling from episodic caches, while others engage more structured, putatively cortex-dependent, representations, which could be updated incrementally, with experience, via an SPE. This distinction might be one reason why hippocampal activity is only occasionally reported in studies of model-based decisions (Simon and Daw, 2011b, Chapters 1 and 2), which instead tend to identify correlates in frontal cortex (Hampton et al., 2006, 2008; Gläscher et al., 2010; Daw et al., 2011).

Much work is needed to clarify the process by which model-based representations are learned. In the experimental studies presented here, we mostly remain agnostic to

the particular process for learning these representations. It may be that multiple types of model use — simulation, inference — coexist (Lengyel and Dayan, 2008), and indeed may themselves each rely on distinct representations of the environment (Doya et al., 2002). Instead, we focus on the *use* of environmental contingencies in behavior, which is the critical signature that underlies all of the potential implementations of model-based control.

In Chapter 1 we isolate a particular kind of contingency, sequential associations between picture stimuli, and identify a unique signature of their representation in hippocampus and use in effecting response behaviors, distinct from striatally encoded action representations. In Chapter 2 we use this signature to show the use of these stimulus-stimulus associations during both simple responses and more complex, deliberative planning for rewards, by a mechanism involving reinstatement of visual content-preferential activity in ventral stream regions. In Chapter 3 we build from the model of this reinstatement activity as a sampling approach to model-based control, and test a specific prediction of this hypothesis: namely, that potentiating particular episodic samples, using associative cues, can impact choices in specific and predictable ways.

Together, these studies advance the notion that the hippocampal memory system supports the representations used to effect model-based control of behavior. I next review previous empirical and theoretical work on the characteristics of model-based control, and compare these results to contemporary understandings of the function of the hippocampal memory system.

The relationship between model-based and model-free control

Numerous features of normal instrumental learning are explained as the result of a shifting — or overlapping — balance of control between model-free and model-based systems (Daw et al., 2005; Chavarriaga et al., 2005). This competition between multiple forms of control — Pavlovian, habitual, and goal-directed — can also explain numerous "pathological" behaviors (Dayan et al., 2006). But the proposed neural substrates of these systems have significant overlap, and questions remain about which neural structures distinctly belong to a given control system.

From a computational perspective, a number of potential mechanisms have been proposed by which model-based value estimates may be used to guide behavior alongside those of the model-free system. For the purposes of discussion, these can be broadly divided into parallel and serial architectures, though the distinction is in practice often far less clear (Botvinick and An, 2008; Bornstein and Daw, 2011; Frank, 2011; Frank and Badre, 2012; Badre and Frank, 2012).

In a purely parallel architecture, the model-based system (or systems; Doya et al. 2002) would produce value estimates or candidate actions at every decision. These estimates, as well as estimates produced by other controllers (e.g., model-free (Daw et al., 2005) or episodic (Lengyel and Dayan, 2008) — though here we treat the episodic controller as another form of model-based control), are then either selected among, or merged in some weighted combination, to effect instrumental behavior (Daw et al., 2005; Keramati et al., 2011; Gläscher et al., 2010; Simon and Daw, 2011a; Daw et al., 2011; Doll et al., 2012, Chapters 1 and 2). A number of proposals exist for parallel operation of model-based and model-free systems, and numerous questions remain unresolved.

For one, if parallel value estimates are produced, but a single behavior is executed, how does each system update its estimates for future behavior? That is, what are the inputs to the prediction error calculation? One possibility is that the update to a system's estimate is weighted in proportion to the degree that system's estimate impacted behavior. Another is that the prediction errors are computed separately as the difference between the reward received and the reward expected for each controller in isolation, and the result reflected in distinct subnetworks (Bornstein and Daw, 2011; Diuk et al., 2013).

This computational distinction among anatomical subunits may bear on empirical findings that different regions of the nucleus accumbens reflect action values for different types of Pavlovian responses; namely, 'consummatory' and 'preparatory' (Konorski, 1967; Wise and Rompre, 1989; Balleine, 2005). Preparatory responses are outcome-insensitive, and thought to be explicitly dependent on the medial region of the accumbens, known as the *accumbens core*. Consummatory responses are those for which the physical manifestations are specific to aspects of the outcome's identity — e.g. licking, salivating — and are proposed to critically depend on the outer layer of the accumbens, the *accumbens shell*. Concordantly, the accumbens shell and core are anatomically connected to starkly different networks of cortical and subcortical regions (Joel and Weiner, 2000). Particularly relevant for the purposes of our discussion is the fact that the accumbens shell is preferentially connected to the hippocampus, and this connection is critical for the sorts of behaviors discussed here, such as contextual conditioning (Ito et al., 2008).

While Pavlovian responding is distinct from instrumental control, this work is highly relevant because the representations required to effect these responses appear to map

well on to formulations of those necessary for model-free and model-based control (Bornstein and Daw, 2011; Bornstein et al., 2011). A direct comparison of the anatomical specificity of preparatory and consummatory responding to that elicited by model-based and model-free control has yet to be carried out, and could further substantiate this important link between theories of Pavlovian and instrumental control (see McDannald et al. (2011) for related work). This is in part due to the difficulty of designing an experiment that thoroughly separates model-based and model-free predictions (but see Chapter 2), but also in part due to limitations of measurement techniques. Namely, simultaneous recording from both the shell and core, or targeted lesions that affect either structure in isolation, are at the edge of feasibility given current technology. Neuroimaging holds promise as a non-invasive technique for simultaneously observing activity in both areas. That no study has yet attempted to observe distinct activity in shell and core may be due in part to limits of the standard resolution and data smoothing methods (but see Choi et al. (2012) for an example of observations using large amounts of data ($n = 1000$) and a comparatively narrower smoothing kernel). It remains a possibility that imaging methods with higher spatial resolution, applied to midbrain structures (De Martino et al., 2013), could yield positive functional distinctions.

The notion of at least some degree of parallelization between the nucleus accumbens portion of control systems sustaining different types of outcome representations is supported by the known anatomy. That is, each subregion exhibits wholly different patterns of connectivity, an anatomical feature preserved in rodent, non-human primate (Joel and Weiner, 2000) and humans (Choi et al., 2012). Again, the functional significance of the stark difference in connectivity between the two subregions of accumbens has not been

thoroughly examined experimentally.

In contrast to the parallel model discussed to this point, a 'serial' design is one in which both the model-based system and direct experience are used to 'train' the model-free cached value system, which is the only place where expected values are compared to experience. Thus, only one reward prediction error is computed, on this model-free value. Though the model-based system may have more or less control over behavior at any given time, the prediction error that updates the model-free system is computed as the difference between experienced reward and the isolated expectations of the model-free system.

The serial approach has received relatively less attention from an empirical perspective, despite a long history of theoretical treatment, most notably incorporated into the DYNA architecture (Sutton, 1991). A critical feature of DYNA is the ability of the model-based system to train the model-free system during offline or rest periods, by replaying experience or sampling from the model (Johnson and Redish, 2005; Foster and Wilson, 2006). (We will return to this replay feature several times throughout our discussion, identifying parallels and suggestive data in models and empirical studies of mechanisms of both choices and memory.) This means that, during rest periods — which may be as short as an inter-trial interval — the model-free value is updated to more closely resemble the model-based value. This feature might partly explain why model-based influences are seen in neuroimaging studies of the nucleus accumbens reward prediction error (Gläscher et al., 2010; Daw et al., 2011; Simon and Daw, 2011b). In Chapter 2, we employ a task in which the model-free and model-based expected values are dramatically distinct — in fact, the model-free system can have no coherent

predictions, as participants have no experience with this particular choice — and find that the RPE signal is wholly explained by model-based expectations (specifically, those computed using a hippocampally-linked contingency model).

Supporting a serial understanding of value estimation, one recent study assessed the criticality of ventral striatum — as a whole — to model-based reinforcement learning, using a paradigm known as 'identity unblocking' (McDannald et al., 2011). They report that the ventral striatum is essential for both types of RL control. Their results support the notion that ventral striatum is a central bottleneck for updating values estimated by both the model-based and model-free system, but do not bear on the question of how those estimates are arbitrated (Bornstein et al., 2011).

Again, though we have treated serial and parallel architectures as distinct for the purposes of exposition, most empirical investigations have considered some hybrid of the two notions in generating testable predictions about factors that affect the prevalence of one system or another over the course of an experiment (Gläscher et al., 2010; Daw et al., 2011; Simon and Daw, 2011a; Otto et al., 2013). The types of mixture models employed in these studies (and in ours, below) do not discriminate between combinations of model-based and model-free influences on a single trial or at different times of an experiment. A study of the factors that influence model-based and model-free control on a single trial level — and, thus, could shed light on whether these estimates are combined in control or alternate dominance — has yet to be conducted.

Sampling as a general heuristic for model evaluation

As I've described above, what constitutes a model is very broadly defined — models can be simple enough to write down in their entirety (e.g., Tic-Tac-Toe) or so com-

plicated that they could not be represented even by a computer the size of the known universe without significant pruning (e.g., Go). Inbetween those extremes, ecologically valid models of many types need to be efficiently evaluated by actors in their everyday environments. Given the reasonable assumption that ecologically valid models are too complex to evaluate in their entirety, and yet need to generate predictions efficiently, Daw et al. (2005) suggest that the heuristic evaluation of models might be a key feature of the tradeoff between control systems. These heuristics could take many forms. For some models, while a complete representation of the world is in theory computable, the time and memory required might be beyond feasible limits. For instance, in playing chess, one could represent the potential next moves and their consequences by a tree structure, which could be fairly compactly represented as a generative model. In theory, one could use this model to evaluate the entire branching tree to select the most optimal move. In practice, however, the order 2^{64} evaluations required would be well beyond the limits of even the most prodigiously talented grandmasters. Heuristics may be employed — for instance, to stop evaluating when a certain pattern is matched, or to look ahead four moves at most, or to only look ahead down paths that match certain rules. In the phrasing of Daw et al. (2005), these heuristics add 'computational noise' — the model *could* be used to evaluate the best possible action of all candidates, but in practice the action it selects is guaranteed to be optimal only up to some degree of uncertainty.

Modeling environments for which the actor can have little specialized knowledge with which to construct compact generative models — or bespoke heuristics to evaluate them — introduces other forms of uncertainty. Rather than chess, an actor might move instead to a completely unfamiliar game, or one whose structure embeds some ir-

reducible uncertainty about action contingencies (e.g., the multi-armed bandit problem). Actors could evaluate these more generalized representations stochastically, by drawing samples of experience. It turns out that, even in this most general case, there are efficient, generally applicable, sampling heuristics that provide reasonable guarantees about the correctness of your model evaluation (Jordan, 1999). These heuristics also provide probabilistic guarantees about the quality of the information that they generate. Thus, an agent that implements these heuristics can have a reasonable estimate of the uncertainty of their evaluations, which would be helpful in implementing the uncertainty-based arbitration of Daw et al. (2005). A rich literature describes sampling-based approaches to complex models. These approaches employ heuristics whose goal is to select actions at a stochastic ratio according to their probability of being optimal in a given situation. Thus, they embed some irreducible stochasticity, by design. These irreducibly stochastic operations have guided the development of new software, programming languages (Goodman et al., 2008) and even hardware (Mansinghka, 2009) which is successfully tackling complex optimization and inference problems, akin to those that humans and animals must solve in order to navigate uncertain environments.

Characteristics of sequential sampling in behavior

The use of a sampling approach to model evaluation and action selection may explain several idiosyncratic features of choice such as nonlinear probability weighting, loss aversion, and hyperbolic discounting (Stewart et al., 2006; Erev et al., 2008a). Patterns of probability distortion in decisions "from experience" — that is, after repeated interaction with an outcome of an action, rather than by description of the space of possible outcomes — may be explained by reliance on relatively small samples of information

(Fox and Hadar, 2006). These sample patterns are inconsistent with the adaptive running average implied by summary statistic representations such as that updated by temporal-difference learning, and may instead be more consistent with replay of small numbers of episodic instances (Vul et al., 2008; Giguere and Love, 2013, Chapters 2 and 3).

Correctly implemented, and in the limit of the number of samples drawn, sampling is the Bayesian optimal solution to inference through uncertainty (Gold and Shadlen, 2002). However, though many studies report that inference behavior is Bayes optimal, this may only be the case when averaging across many observers. Individuals themselves may rely on only single samples, which could underlie a wide variety of suboptimal or idiosyncratic behavior (Vul et al., 2008).

Episodic sampling as a distinct form of control

This approach of sampling from episodes of past experience is distinct enough from traversal of structured representations that it has been posited as a third form of control, optimally useful in environments with which the actor has little experience (Lengyel and Dayan, 2008). A key distinction, with potentially deep parallels in cognitive neuroscience, is that between sampling from episodes of experience and sampling from *forward models* of the environment. In the former, individual experiences are selected according to some sampling heuristic, and the action and outcome in them are used to guide current behavior — e.g., if I sample a trial on which I chose option A and gained \$5, I'm more likely to repeat that action; if I sample a trial on which I chose option B and lost \$5, I'm more likely to choose the other action. In the latter, samples might take the form of probabilistic outcomes of a generative model — for instance, if I hear that the grass outside is wet, then depending on my environment and the likelihood of

certain contextual features, I might sometimes decide that it is wet because of rain, or I might sometimes decide that it is wet because the sprinklers were turned on; critically, in neither case do I necessarily need to recall a particular instance of the grass being wet. In either representation, discrete actions are suggested as candidates for current behavior, based on an internal representation. To the casual observer, actors using either of these internal representations might behave indistinguishably in quotidian settings.

For the studies in Chapters 1 and 2, we make no claims about which of these two types of representation might be guiding behavior — only that a representation exists, and that its use corresponds to activity in the hippocampal memory system and connected cortical regions. In the experiments of Chapter 3, we use an experimental manipulation that favors specific episodes, more strongly identifying behavior as being driven by samples drawn from an episodic cache.

Importantly, the second type of representation can yield far more flexible, context-dependent behaviors, and may be supported by a distinct set of neural structures from those involved in episodic cache. Generative, semanticized representations require extensive experience to develop and encode efficiently (McClelland et al., 1995). The processes by which they develop are as yet poorly understood, but might be related to mechanisms of memory consolidation (Carr et al., 2011), especially when this consolidation yields changed representations that can impact future behavior (Tambini et al., 2010). The neural mechanisms by which these representations develop (Kumaran et al., 2009), and how their use might be arbitrated against the use of episodic caches, is an exciting topic for future research.

It is worth noting (and we later return to the idea) that the process of making deci-

sions by sampling from candidates for ensuing experience bears similarity to the phenomenon of "preplay" famously associated with hippocampal activity preceding certain types of decisions (Ferbinteanu et al., 2003; Johnson and Redish, 2007). The proposal that noisy evidence from memory informs decisions via a sampling process dates back to at least the mnemonic accumulator hypothesis (Ratcliff, 1978), an early application of the drift-diffusion models that have been widely applied in the study of perceptual decisions. However, unlike perceptual information, there is not strong agreement on a method for quantifying memory information. The approach developed in Chapter 1 offers a way to quantitatively measure a particular kind of memory — that for temporal order – by expressing it in the same units as noisy perceptual information, as a multinomial probability distribution. This tool might be of value in studies of how hippocampally-linked information impacts a wide variety of memory-guided behaviors, from perception to action. In Chapter 2, we use this technique to observe the use of memories in effecting model-based control of decisions. This study provides a computational interpretation of the role of hippocampus in instrumental control.

We now review previous empirical and theoretical work describing how the hippocampus contributes to a wide variety of behaviors.

The hippocampal memory system in behavioral control

It is well-established that the hippocampus supports learning functions parallel and complementary to those supported by the striatum (Packard et al., 1989). These two types of learning — alternately referred to as declarative versus procedural or episodic versus habitual — are traditionally thought of as in competition for control of behavior

(Knowlton et al., 1996; Poldrack et al., 2001).

The hippocampus and related areas of the Medial Temporal Lobe (MTL) — such as the entorhinal and parahippocampal cortices — together comprise the hippocampal memory system. This network of structures is strongly identified with key roles in episodic memory and representations of space. The encoding of these memories is characterized by the rapid, single-shot, binding of associations tying together many constituent elements. Here I review some of the key work that describes the representations supported by the hippocampal memory system. I focus on how these representations are learned, how they change with experience (and rest), and how they might be deployed in support of behavior.

Representations of the environment in the hippocampal memory system

The hippocampal memory system is famously the critical site of multiple representations of environmental structure — in particular, the celebrated 'place cells' and 'grid cells' (O'Keefe and Nadel, 1978; Fyhn et al., 2004; Hafting et al., 2005).

Place cells encode specific regions within a given environment, with scale and density of the representation potentially corresponding to the behavioral relevance of those regions (Zinyuk et al., 2000). They fire at a given location within a given space, and as rodents travel along the space, the cells can be seen to fire in sequence along with the animal's travels. After multiple traversals of a given space, Hebbian processes link sequences of well-traveled paths into *cell assemblies*, which together can be thought of as encoding trajectories through space — and, critically, implies that the action of one member of the cell assembly increases the probability that further members of that cell assembly will also become active. As I discuss below, this chain of associated locations

may serve as a useful substrate for decision-making. This representation 'remaps' in new environments: a given cell may dramatically shift its response properties between even adjacent rooms of a maze (Knierim et al., 1995). The place cell representation has been shown to guide navigation by supporting the creation of a representation of a goal location (Burgess et al., 2000). These rich internal models support a wide variety of ecologically observed behaviors, such as path integration (Etienne et al., 1998; Etienne and Jeffery, 2004). Disrupting these representations — either via lesions (Pearce et al., 1998) or targeted interference of specific windows of spiking activity known to be critical for their maintenance (Jadhav et al., 2012) — severely impairs the ability of animals to navigate without external cues.

Grid cells provide a different, putatively complementary representation (Hafting et al., 2005). They fire at regular, periodic intervals within a given space, with response fields that tend to fall along a triangular 'grid'. Whereas place cells fire at particular locations in an environment, grid cells overlay these landmark-like locations with a metric of space.

Both representations are thought to be key for the practical use of spatial knowledge. Computational studies suggest that simultaneously learning contingencies at multiple scales of time and space has been shown to greatly improve performance, at little additional computational cost (Singh, 1992; Dayan and Hinton, 1993). Work in machine learning has shown that a rich, flexible forward model of the environment, comparable to hippocampal place cells, can be learned during exploration of an unfamiliar environment by integrating multisensory information using simple Hebbian processes (Arleo et al., 2004). Learning how to act optimally in noisy, uncertain environments

is a famously difficult problem to solve efficiently, leading to the use of heuristics and shortcuts, whose performance is strengthened by the availability of rich representations of associative structure in the environment being navigated (Littman et al., 1995). Further, the co-existance of multiple such models, each tuned to enable different types of behavior, might be crucial to flexible and efficient navigation (Narendra et al., 1997). Supporting the idea that these representations exist to guide behavior — rather than to map space, per se — the metrics that define their scale may be modulated by the geodesic distance between points — e.g., barriers would introduce distortions into the representations — rather than euclidean distances (Gustafson and Daw, 2011).

I next discuss recent findings about how these representations, in particular place cells, operate during spatial decision-making.

Functions of the hippocampus in spatial decisions

The hippocampus has long been understood to be critical for normal spatial navigation in rodents (Tolman, 1948). Though place cells are widely thought to be important for this function, the precise mapping of place cell firing to observed behavior remains a topic of significant ongoing investigation (Burgess et al., 2007; Hasselmo and Eichenbaum, 2005; Lisman and Grace, 2005; McNaughton et al., 2006; Moser et al., 2008).

The hippocampus is necessary for performance in a spatial alternation task with delays between trials (Ainge et al., 2007). This observation concords with a proposed role for the hippocampus in 'temporal credit assignment' — the ability to track back a received reward to the instrumental cause that should be reinforced (Foster and Wilson, 2006). The activity that inspired Foster and Wilson (2006) to suggest a role for hippocampus in temporal credit assignment is called 'reverse replay'. During periods

of rest or pauses in activity after a navigation trial, the researchers observed firing of place cells *in reverse order* — that is, the cell assemblies would be triggered beginning at the goal location, and follow back to the start. This firing pattern conceivably strengthens the connections among members of the assembly in the reverse direction, making them usable for goal-directed decisions — decisions that aim to achieve a particular outcome, and must choose a path that leads to the selected goal. Concretely, after reverse replays, the activation of a goal representation — for instance, during the act of deliberation about a navigation decision — should increase the firing probability for backward-chained members of the cell assembly representing the trajectory that lead to that goal representation.

Indeed, evidence is emerging that the hippocampus subserves a critical role in goal-directed choice, especially during early learning of a choice environment, and replay events appear to be necessary for effecting hippocampally-driven choice behavior (Ego-Stengel and Wilson, 2010). Replay activity has been observed both following outcome receipt, by Foster and Wilson (2006) above, and *preceding* choice (Ferbinteanu et al., 2003; Johnson and Redish, 2007). These replay events can reflect trajectories beginning at either the current (Johnson and Redish, 2007) or remote (Davidson et al., 2009; Gupta et al., 2010) locations — importantly, they are not a simple function of current sensory input (Karlsson and Frank, 2009) — and cell firing can reflect a forward or backward traversal of the experienced path (Gupta et al., 2010). Several such events occur during pauses prior to decisions — suggestive of a neural equivalent to the observation by Tolman (1948) of 'vicarious trial-and-error' — and can encode the content of many trajectories during a single epoch. When preceding correct decisions during early learn-

ing of a spatial alternation task, this activity is both more structured and evidences a structure that is biased towards correct trajectories (Singer et al., 2013).

These replay events are thought to be carried by periodic "sharp wave ripples" (SWRs) (Buzsáki, 1986). Critically, these ripple events are common features of regular hippocampal activity, occuring during both sleep (O'Neill et al., 2008) and periods of waking rest (Ferbinteanu et al., 2003; Foster and Wilson, 2006). SWRs may play a role in reinstating — and, thus, orchestrating further coactivation of — cortical patterns associated with the original experience (O'Neill et al., 2010). Cell coactivation patterns during sleep and resting ripples reflect those observed during waking behavior, with proportionally greater coactivation observed in 'offline' periods for cell assemblies that fired together more often while awake (O'Neill et al., 2008). These data support the hypothesis that apparently spontaneous offline activity is 'replaying' past experience.

Altogether, these data suggest that ripple events, both online and offline, may serve a critical function in spatial decision making, one that might be well-described by the sampling algorithms described in the previous section. I next discuss work suggesting that these findings generalize to domains other than spatial navigation, with particular focus on the sequential associations that are relevant to the studies presented in Chapters 1 and 2.

Hippocampus in sequential learning

I have to this point reviewed the function of the hippocampus primarily through the lens of spatial learning. While there may in fact be key differences in how hippocampal representations are used for spatial navigation in rodents and in humans (Dede et al., 2013), the emphasis on spatial navigation in part reflects the fact that a great deal of

hippocampal work has been performed in rodents, for whom spatial navigation tasks are well-developed and largely guide research questions. It is widely believed that the bulk of the understanding gained about hippocampal function in representing space and using these representations to effect behavior also translates to non-spatial domains (Eichenbaum, 2004). A less well-known, but still extensive, set of studies explores the role of the hippocampus in binding sequential associations — stimuli linked across time instead of (or in addition to) space. This role is of particular relevance to the studies I present in Chapters 1 and 2.

The hippocampus has long been implicated in binding together elements across time as well as space (Dusek and Eichenbaum, 1997; Shohamy and Wagner, 2008; Wimmer et al., 2012). Recently, a number of studies have explored the role of the hippocampus in sequence learning, behavior previously presumed to be largely procedural in nature (Keele et al., 2003; Strange et al., 2005; Harrison et al., 2006; Kumaran and Maguire, 2006; Turk-Browne et al., 2010, 2012). An overview of this work yields common observations including: that overall hippocampal activity is greatest when sequential structure is most uncertain, or when elements of the sequence are recalled in greater detail, that hippocampally-linked activity encodes predictive relationships among sequence elements, and that — contrary to long-held presumptions — it may correspond to either explicit or implicit awareness of sequential structure. These data are broadly consistent with the notion that hippocampal cell assemblies form over repeated experience with sequentially experienced locations, and that they are shaped by their presumed relevance to behavior.

Recent data obtained using lesions in humans and animals further supports the ne-

cessity of hippocampus for linking across time. As the temporal context linking action and outcome extends, the hippocampus becomes critical for correct learning from feedback (Foerde et al., 2013). Hippocampal lesions greatly impair learning on a spatial alternation task (Kim and Frank, 2009), consistent with a role in the acquisition and integration of sequential contingencies over repeated experience. Paradoxically, however, hippocampal lesions can *improve* performance — expressed as an increase in the behaviorally expressed learning rate — on reversal learning tasks (Eckart et al., 2012; Will et al., 2013). This observation is consistent with our results in Chapters 1 and 2 and those of others (Komorowski et al., 2009), which describe hippocampally-expressed behavioral control as drawing on experiences over a long timescale of memory. Incorporating episodes from the long past may *impair* effective responding on tasks where contingencies change suddenly or rapidly, such as probabilistic reversal learning tasks. We expand on this idea in in the study presented in Chapter 3, and exploit it to affect choices between options with shifting reward contingencies.

I next discuss how hippocampal representations change over time, and review suggestions about how this process may also impact behavioral control.

Rest, consolidation, and prediction

Tying together the preceding sections is the suggestion that the hippocampus subserves a general function of binding together associated experiences in a manner that best supports future behavior. The support of future behavior by anticipatory activation hippocampal cell assemblies has been termed 'prediction', perhaps reflecting a bias towards explicit memory metaphors in describing hippocampal function. Critically, the proposal that prediction is a key function of the hippocampal memory system implies that the

representations it encodes are shaped in a manner that can most efficiently be applied to anticipated future behavior.

The idea that hippocampal representations are geared towards the adaptive support of behavior has a long history in psychology (Tolman, 1948) and cognitive neuroscience (Cohen and Eichenbaum, 1993). That consolidation produces more adaptive representations, via repeated reinstatement, is a key feature of the predominant computational model of the neural representation of declarative memories (McClelland et al., 1995). Reviewing the literature on memory with an eye towards its role in prediction, Buckner (2010) defines the adaptiveness of the hippocampal memory system in these terms: "the capture of associations that define event sequences is adaptive *because* these sequences can be reassembled into novel combinations that anticipate and simulate future events" (emphasis added).

This perspective is useful when considering the activity of the hippocampus when there is *no* active behavior being performed: when the animal is resting, or even asleep. A key feature of representations supported by the hippocampal memory system is that they undergo *consolidation*, or the transition of memories to a cortically-dependent, and hippocampally-independent, phase (McClelland et al., 1995). Though they persist and critically support long-term recall of information and episodes, consolidated memories are not simply copies of the original hippocampal cell assemblies. Instead, they are "recoded, chunked, or otherwise derived forms of the original experiences" (Buckner, 2010). A reasonable suggestion is that this transformation occurs in a manner shaped by the suitability of the derived representation for behavior. Lengyel and Dayan (2008) suggest that consolidation might improve performance of the episodic controller, even

in low-data conditions, by "eliminating unfortunate sample trajectories". And this suggestion is supported by the fact that putatively consolidation-containing rest improves behavioral performance in a wide variety of cognitive tasks (Durrant et al., 2011; Fischer et al., 2002; Nitsche et al., 2010; Schacter et al., 2011; Press et al., 2005; Walker and Stickgold, 2004).

Replay activity is a potential mechanism for consolidation as well as retrieval (Maquet, 2001; Walker and Stickgold, 2006; Carr et al., 2011). The firing patterns of sleep replay events occur in the same temporal order as the original experiences (Skaggs et al., 1996; Lee and Wilson, 2002), suggesting that they preserve and strengthen the experienced associations. During sleep, replay events occur not just in the hippocampus, but also in associated cortical regions (Euston et al., 2007), while thalamic activity is markedly diminished, perhaps to minimize sensory input (Logothetis et al., 2012), which has been shown to influence the content of awake replay (Carr et al., 2011).

During resting states — both awake rest and sleep — a broad, interconnected range of brain structures are preferentially active, above levels seen during performance of many tasks. This network — known variously as the "default network" or the front-temporal resting-state memory network — has been observed in humans (Raichle et al., 2001), non-human primates (Vincent et al., 2007), and rodents (Desai et al., 2011).

Despite its name, the regions of the resting state network are not exclusively coactive during offline states. Indeed, functional hypotheses abound for multiple behaviors to which the network might be critical. Though these behaviors have been identified in several tasks and are known by many names — prospective simulation, mindwandering, constructive episodic memory, and goal-directed planning — they have in common

the recombination of past experience to subserve novel demands (Addis et al., 2007; Buckner, 2010; Gilbert et al., 2013). This is precisely the key feature of model-based control. Thus, a unifying explanation is that both online and offline activity in the network involves simulating or sampling from past experience, and both have the common consequence of encouraging the formation of long-term cortical representations of this experience.

Linking these observations with the computational ideas reviewed in the previous section suggests a role for consolidation in the transfer between multiple forms of model-based control. Specifically, the repeated sampling of episodic memories during offline states, leading to the construction of recoded, compressed, semanticized representations linking multiple cortical areas, might correspond to the transition from episodic control to more structured representations attuned to particular behaviors or types of evaluation. One prediction of this model is that previously hippocampally-dependent goal-directed behaviors (like those studied in Chapters 1 and 2) would preferentially engage cortical regions after a period of rest. Another prediction is that activity during rest would reflect accelerated learning of those representations by cortical regions, using hippocampal replay as input instead of sensory experience. Therefore, signatures of semantic structure learning, such as State Prediction Errors (SPEs), might be observed in regions subserving the relevant consolidated representations, and might be contemporaneous with neural signatures of hippocampal model engagement (as in Chapter 1 and 2).

Though the field is just at the very beginning of understanding the mechanisms and behavioral relevance of offline activity, it may prove fruitful to analyze this activity in terms of model-based learning and control.

Interim summary

In this introduction I have described a class of decision-making algorithms known as model-based reinforcement learning. I have outlined recent theoretical and experimental work on a number of approaches to learning and making decisions using learned representations of environmental contingencies. I then reviewed experimental literature on the functions of the hippocampus in decision making. I argue that this work has revealed that memory replay events, carried by sharp-wave ripples, serve as a substrate for sampling-guided decisions of the sort proposed to heuristically evaluate models in model-based RL.

This compendium of observations suggests that the hippocampal memory system is an important factor in the normal execution of a wide range of behaviors, from simple responses to complex, goal-directed choice. But does this multitude of effects have a common cause? Is there a unifying, parsimonious description of the role of the hippocampal memory system in effecting multiple forms of instrumental control?

In computational terms, the hippocampal memory system is central to the 'cognitive map' or 'world model' that supports model-based decision making. In summary, the key distinguishing feature of model-based control is representational — behaviors supported by the hippocampus are those critically dependent on the ability to flexibly chain together previously learned information in the service of new goals. The representations necessary to effect this kind of control have long been associated with the prefrontal cortex (Miller and Cohen, 2001; Daw et al., 2005). But these proposals have presumed well-formed, executable or navigable representations of the type understood to be encoded by association cortex. I have argued above that the hippocampal memory system

deserves further consideration as a candidate for model-based control of a different kind. I proposed that the role of the hippocampal memory system in instrumental control is to provide a substrate for generalizing across multiple contextual factors that might influence behavior. Specifically, that the types of representations formed by hippocampus — episodic memories — can serve as 'samples' for approximate inference of a developing or incompletely structured world model, and that these samples are drawn for the purpose of identifying past contexts similar to the current one. Observations of hippocampal 'replay' and 'preplay' activity before choice bear signatures of sample-based approaches to decisions. That this sampling is ongoing, even long after learning even of simple tasks has reached asymptote, suggests that though value estimates emerging from samples appear to drive decisions primarily in situations of low information (or low information quality), they may continue to impact decisions long afterwards. This continuing influence of episodic samples could serve as an explanation for occasional exploratory behavior or other idiosyncratic patterns in choice (Stewart et al., 2006; Erev et al., 2008a).

In the remainder of this document, I present three empirical studies that address key questions about the nature and use of the hippocampal memory system in model-based control.

In Chapter 1, we develop a signature of the influence of hippocampal representations on behavior. Importantly, we are able to seperately identify the effects of hippocampal predictions in simple motor responses, and use this as a tool to isolate behavioral and neural activity related to hippocampally-learned representations.

In Chapter 2, we use the signature developed in Chapter 1 to identify the exclusive

use of the hippocampal representation in effecting model-based control. We modify the sequential learning task from Chapter 1, adding choice trials on which we ask participants to leverage their knowledge of associations to make trial-unique, goal-directed choices. We develop a model of how the hippocampal representation is deployed to effect choice, and test neural predictions of that model. The model posits that past instances of sequential stimulus-stimulus associations are reinstated during deliberation, accumulated as samples of evidence towards a decision. We find that patterns of activity in ventral cortical regions reflect the reinstated stimulus categories during the formation of both simple responses and complex, goal-directed choices.

In Chapter 3, we explore the sampling hypothesis in more detail. Specifically, we assess the ability of a sampling approach to fit choice behavior in a canonical RL task, choosing between slot machines of continuously changing payoffs. Critically, we demonstrate a sampling model that generates behavior with a strikingly similar pattern of dependency on past experience to that of a temporal-difference model, but with a pattern of choice probabilities that proves a superior fit to choices. As a final experiment, we test a specific prediction of the sampling model: namely, that particular episodes can impact individual choices. We employ a manipulation designed to privilege particular episodes of trial experience from the relatively distant past over those more temporally recent, and show that these episodes do in fact impact subsequent choices in specific, predictable ways.

Together, these studies support the claim that the representations of the hippocampal memory system are a critical component of model-based control of behavior.