From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Neuroscience

2019

Meta-Reinforcement Learning with Episodic Recall: An Integrative Theory of Reward-Driven Learning

Ritter, Samuel

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01dz010s84f

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

Meta-Reinforcement Learning with Episodic Recall: An Integrative Theory of Reward-Driven Learning

Samuel Ritter

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Neuroscience
Adviser: Matthew Botvinick

January 2019

© Copyright by Samuel Ritter, 2019.

All rights reserved.

Abstract

Research on reward-driven learning has produced and substantiated theories of model-free and model-based reinforcement learning (RL), which respectively explain how humans and animals learn reflexive habits and build prospective plans. A highly developed line of work has unearthed the role of striatal dopamine in model-free learning, while the prefrontal cortex (PFC) appears to critically subserve model-based learning. The recent theory of meta-reinforcement learning (meta-RL) explained a wide array of findings by positing that the model-free dopaminergic reward prediction error trains the recurrent prefrontal network to execute arbitrary RL algorithms – including model-based RL – in its activations.

In parallel, a nascent understanding of a third reinforcement learning system is emerging: a non-parametric system that stores memory traces of individual experiences rather than aggregate statistics. Research on such *episodic learning* has revealed its unmistakeable traces in human behavior, developed theory to articulate algorithms underlying that behavior, and pursued the contention that the hippocampus is centrally involved. These developments lead to a set of open questions about (1) how the neural mechanisms of episodic learning relate to those underlying incremental model-free and model-based learning and (2) how the brain arbitrates among the contributions of this abundance of valuation strategies.

This thesis extends meta-RL to provide an account for episodic learning, incremental learning, and the coordination between them. In this theory of episodic meta-RL (EMRL), episodic memory reinstates activations in the prefrontal network based on contextual similarity, after passing them through a learned gating mechanism (Chapters 1 and 2). In simulation, EMRL can solve episodic contextual water maze navigation problems and episodic contextual bandit problems, including those with Omniglot class contexts and others with compositional structure (Chapter 3). Further, EMRL reproduces episodic model-based RL and its coordination with incremental

model-based RL on the episodic two-step task (Vikbladh et al., 2017; Chapter 4). Chapter 5 discusses more biologically detailed extensions to EMRL, and Chapter 6 analyzes EMRL with respect to a set of recent empirical findings. Chapter 7 discusses EMRL in the context of various topics in neuroscience.

Acknowledgements

I'd like to thank my advisor, Matt Botvinick, who gave me freedom, guidance, and space to work through unexpected challenges. My PhD cohort at PNI: Matt Panichello; Gecia Bravo-Hermsdorff; Diana Liao; Luis Piloto; Alex Piet, who taught me more about math and neuroscience in our first year than all of our lectures combined; and Kim Stachenfeld, who has been an incredible friend and roommate on both sides of the Atlantic. My collaborators at Princeton and DeepMind especially Francisco Pereira, David Barrett, Jane Wang, and Zeb Kurth-Nelson. My parents, who supported me through all of this and my grandparents, both those who are still with me and those who have passed on. And finally my partner in life and sometimes in science, Cotie Long.

Contents

Abstract iii

Acknowledgements v

List of Figures x

1 Introduction and Précis **1**

1.1 Incorporating the Environment's Structure 3

1.1.1 Structure, Novelty, and Reoccurrence 5

1.2 An Integrative Account of Episodic Learning 7

1.3 Testing the Theory with the Episodic Two-Step Task 10

1.4 Recall of Rapid Inferences for Deep RL Agents 13

1.5 Extensions, Connections and Implications 17

1.5.1 Episodic Meta-RL in the Brain 17

1.5.2 Empirical Connections 21

1.5.3 Relationship with Other Theories and Topics 23

1.6 Summary and Roadmap to the Thesis 24

2 Episodic Meta-Reinforcement Learning **26**

2.1 Background and Preliminaries 26

2.1.1 Dopamine and Model-Free RL 26

2.1.2 Prefrontal Cortex and Model-Based RL 28

2.1.3 Hippocampus and Episodic Learning 29

2.1.4 Outstanding Questions 30
2.2 Meta-Reinforcement Learning 31
2.2.1 Meta-RL Implementation for Simulations 32
2.2.2 Meta-RL Simulation Results 35
2.2.3 Meta-RL Summary 37
2.3 Meta-RL with Episodic Recall 38
2.3.1 Motivation 38
2.3.2 Formal Environment Model 39
2.3.3 Architectural Principles and Design 41

3 The Efficacy of Episodic Meta-RL 48
3.1 Experiment 1: Barcode Bandits 48
3.2 Experiment 2: Omniglot Bandits 50
3.3 Experiment 3: Compositional Bandits 53
3.4 Experiment 4: Contextual Water Maze 55
3.5 Discussion 56

4 The Episodic Two-Step Task 62
4.1 Introduction 62
4.2 Task 65
4.2.1 Learning Algorithms 67
4.3 Methods and Results 68
4.3.1 Analysis of Reinstatement Gate Activations 70
4.4 Discussion 73

5 Episodic Meta-RL in the Brain 77
5.1 Prefrontal Working Memory and the LSTM 77
5.1.1 Backpropagation 83
5.2 Hippocampal Episodic Memory and the Differential Neural Dictionary 85

5.2.1 Attractor Networks 85
5.2.2 The Hebb-Marr Model of the Hippocampus 86
5.2.3 EMRL and Hebb-Marr 89
5.3 Gated Reinstatement: Connecting Working and Episodic Memory . . 91
5.3.1 Modeling Directions: Implementational-Level Analogues to EMRL 93

6 Empirical Connections **96**
6.1 "Reminders of past choices bias decisions for reward in humans" – Bornstein et al. 97
6.1.1 Experiment 1 97
6.1.2 Experiment 2 99
6.2 "Reinstated episodic context guides sampling-based decisions for reward" – Bornstein and Norman 101
6.3 "Episodic Memory Encoding Interferes with Reward Learning and Decreases Striatal Prediction Errors" – Wimmer et al. 102
6.4 "Memory States Influence Value-Based Decisions" – Duncan and Shohamy 104
6.5 "Refresh my memory: Episodic memory reinstatements intrude on working memory maintenance" – Hoskin et al. 106
6.6 "Multiple Processes in Prospective Memory Retrieval: Factors Determining Monitoring Versus Spontaneous Retrieval" – Einstein et al. . . 109
6.7 General Discussion 111

7 Broader Connections **114**
7.0.1 Temporal Context Model 114
7.0.2 Serial Recall 115
7.0.3 Rational Accounts of Memory 116

7.0.4 Gated Working Memory Models 117
7.0.5 Planning and Preplay 117
7.0.6 Replay, Consolidation, and CLS 118
7.0.7 REMERGE and Flexible, Rapid, Relational Generalization . . 118
7.0.8 Novelty Detection, Storage, and the Hippocampal-VTA Loop . 119
7.0.9 Grid cells 120
7.0.10 Task Sets 121
7.0.11 Extinction and Renewal 123
7.0.12 Statistical Learning over Individuals and Events 125

Bibliography **127**

List of Figures

2.1 Depiction of the meta-RL agent architecture. A recurrent neural network modeling the prefrontal network receives as input a state observation, its previous action, and the reward received on the previous timestep. The network outputs an action and an estimate of the value of the current state. The network is optimized using gradients from an actor-critic loss function which computes a function of the difference between expected and received reward (Equation 2.14) analogous to the reward prediction error associated with dopamine. Reprinted from Wang et al. (2018). 32

2.2 Model architecture and environment structure. Tasks, such as multi-armed bandits, are presented sequentially to the agent along with identifiable contexts, which are depicted here as colored bars. On each time step, the agent reads from long-term memory (DND) to retrieve cell states, which are reinstated through the multiplicative reinstatement gate (r-gate). At the end of each task, it writes its cell state to the DND. 39

2.3 The differentiable neural dictionary (DND) with state-action value estimates stored in the values column (Blundell et al., 2016). The state observation s is embedded by a convolutional neural network to produce a query key h. The values paired with the nearest neighbors h_i in the DND are aggregated to produce a value estimate for state s. Reprinted from Blundell et al. (2016). 42

2.4 (Top panel) The recurrent working memory of EMRL, receiving state observation, previous reward, and previous action as input and outputting a policy on each timestep. (Middle Panel) EMRL's storage process. The cell state of the LSTM is appended to the DND, along with an embedding of the state ("cue"). (Bottom Panel) EMRL's retrieval process. A similarity based lookup between the current state embedding and the stored state embeddings is performed. The cell states associated with the most similar stored state embeddings are retrieved and reinstated to the ongoing working memory process (see Figure 2.5 for details of reinstatement). 46

2.5 A schematic of the LSTM with episodic memory access via reinstatement gate. In gray, the standard LSTM architecture. In bold, the proposed episodic memory and reinstatement pathways. 47

3.1 (Top) Barcode and (Bottom) Omniglot bandit training curves, averaged over 3 runs. 51

3.2 (Top) Cumulative regret (expected loss in cumulative reward due to selecting suboptimal arms) for EMRL agent on the Omniglot bandits task, averaged over evaluation episodes. Regret decreases as the number of exposures increases. The agent performs similarly to Gittins indices, Upper Confidence Bounds (UCB), and Thompson sampling on its first exposure, and outperforms them on subsequent exposures. (Bottom) MRL with no episodic memory performs on par with standard algorithms on Omniglot bandits, with no decrease in regret with more exposures. 58

3.3 (Top) Training and (Bottom) regret curves for compositional bandits. Training curves averaged over 3 runs, regret curves averaged over evaluation episodes. 59

3.4 (Top) Training curve averaged over 5 runs. The brown line indicates the maximum reward achievable without episodic memory. (Bottom) Number of steps before the EMRL agent reaches the goal after respawning, binned by number of previous exposures to the current context, data from evaluation episodes. After only one exposure, the agent can perform close to optimally. 60

3.5 (Top) Sample trajectories captured during evaluation for episodes with the same starting position and end goal during the first exposure to a context. (Bottom) Same as c, but for the second exposure. Early steps are colored more red; later steps more green. 61

4.1 Contextual two-step task modeled after Daw et al. (2011) and Vikbladh et al. (2017). (Top) Two trial types are shown: uncued and cued. All trials start in state s_0 at the first stage, at which point agents are presented with either a "no-cue" stimulus or are cued with a second-stage stimulus seen on a previous trial. Transition probabilities after taking actions a_1 or a_2 are depicted in the graph. On uncued trials, s_1 and s_2 result in Bernoulli rewards with probabilities r_a and r_b. On cued trials, transitioning into the same state as in the trial k during which the cue was first presented results in receiving the same reward received on trial k, r^*_a. (Bottom) Trials within an episode are split into 4 blocks, with block 3 consisting of cued trials which are cued with stimuli from block 1, and block 4 cued from block 2. 66

4.2 EMRL achieves more reward than MRL. Average reward obtained by MRL and EMRL on cued and uncued trials. EMRL earns more reward than MRL on cued trials, suggesting that EMRL can use its episodic memory to exploit the task's episodic structure. For comparison, a random policy achieves 0.5 reward on this task. Bar heights indicate means over 10 trained agents with different random seeds. Error bars indicate the standard error of the mean over the 10 trained agents. . 71

4.3 Parameter estimates in a model with weighted contributions of four decision systems. The estimates provide evidence that incremental and episodic model-based valuations contribute to EMRL's behavior. The model was fit to EMRL's actions separately on cued and uncued trials. Each gray dot denotes the model fit parameter for one of the 10 trained agents. ** denotes significant difference from 0 at p¡0.005 by binomial sign test. 72

4.4 EMRL exhibits both incremental and episodic model-based behavior. Bar heights indicate means over 10 trained agents with different random seeds. Error bars indicate the standard error of the mean over the 10 trained agents. (Top) Proportion of uncued trials in which EMRL repeated the action it took on the previous trial ($t-1$), split by whether it received reward on $t-1$ and whether the transition on $t-1$ was common. The interaction between those two factors is a sign of model-based learning[1]. (Middle) Same as top, but for cued trials. (Bottom) Same as top, but split by whether EMRL received reward on the past trial k when the cue was first encountered and whether the transition on trial k was common. 75

4.5 (a) Timecourse of the mean values of the reinstatement gate averaged over 500 episodes, split by stage of the trial. (b) Mean values of the reinstatement gate on cued trials on which the agent selected the optimal action (correct) and on cued trials on which it selected the opposite action (incorrect), averaged over all units. 76

5.1 Diagram of the basal-ganglia–thalamic–cortical gating mechanism theorized to support working memory updating. Reprinted from Frank et al. (2001). 80

5.2 Summary diagram of input, output, and reallocation gates in the brain, useful for drawing an analogy from the LSTM's components to potential anatomical substrates. See main text for full description and connection to LSTM. 83

5.3 (Top) Reprinted from (O'Reilly and Munakata, 2000). (Bottom) Reprinted from (O'reilly and McClelland, 1994). 88

Chapter 1

Introduction and Précis

How do we learn from the results of our decisions to make better decisions in the future? Building on psychology's early foundations (e.g., Thorndike, 1911; Tolman, 1948), theories of reinforcement learning (RL) produced breathtaking advances toward answering this question. First, the core theory of RL described the abstract computational challenges inherent in learning from decisions (Sutton and Barto, 1998), then decades of work developed algorithms and exposed the neural implementations underlying this type of learning in humans (reviewed in Dolan and Dayan, 2013). Moreover, these algorithmic theories double as algorithms *per se*, variants of which have revolutionized the landscape of engineering approaches in machine learning (e.g. Mnih et al., 2015; Silver et al., 2016). However, RL faces fundamental challenges as a computational theory, algorithmic/implementational theory, and as an engineering discipline.

First, consider RL's well-known, compact, and widely applicable computational framework: the environment consists of states, action-conditional probabilities of transitioning among them, and state-reward probabilities; the agent's goal is to max-

A note on the organization of this thesis: this introductory chapter is intended to function as a self-contained executive summary, conveying in a tightly abridged format most of the key ideas and results of the thesis. Readers may use this introduction as a portal to the rest of the thesis, following the references to later chapters when interested in greater detail.

imize reward. The sweeping generality of this framework is one of the strengths that led to RL's widespread success; however, it is also a source of critical weakness as a psychological theory. Because this environment description is so general, it has little to say about the structure of the natural world, and therefore it is not possible to derive from it the inductive biases that humans and animals exhibit after experience in natural environments. To create a theory that provides this kind of insight, it will be necessary to build on RL's computational framework — to define models of the environment with states, actions, transitions, and rewards that reflect essential computational properties of the natural world. Psychological research in RL has thus far selected these quantities not with the goal of modeling ecology, but instead with the goal of isolating aspects of behavior or cognition. This focus on tightly controlled experiments that can practically be applied to humans and animals in the lab has led to groundbreaking discoveries while leaving an impoverished computational theory.

Consider next RL's most prominent contributions to algorithmic and implementational level theory: temporal difference and model-based learning. Both have had considerable success in explaining behavioral and neural findings (Daw et al., 2011, 2005; Gläscher et al., 2010; Schultz et al., 1997; Bayer and Glimcher, 2005; Cohen et al., 2012; Hare et al., 2008), however, neither explain how humans and animals dramatically update their behavior after a single experience (Lee et al., 2015). Further, human behavior shows signatures of episodic learning, which these dominant RL theories fail to account for (Vikbladh et al., 2017; Bornstein et al., 2017; Bornstein and Norman, 2017; Wimmer et al., 2014; Wimmer and Buechel, 2016; Duncan and Shohamy, 2016). A relatively recent development in RL theory has taken steps toward filling this gap, postulating that the brain may carry out kernel-based episodic sampling over past experiences in order to compute value estimates in novel states (Gershman and Daw, 2017; Lengyel and Dayan, 2008). A remaining problem is that the explanations of model-based, model-free, and episodic learning remain disjoint,

raising mechanistic questions about how the implementations of these algorithms relate to one another and functional questions about how these algorithms are arbitrated among.

Finally, as an engineering approach RL has achieved ground-breaking advances, but learning is generally extremely sample inefficient, and a consensus holds that this inefficiency is due to lack of prior knowledge (Lake et al., 2017; Finn et al., 2017). Recent approaches to RL that incorporate prior knowledge (Finn et al., 2017; Duan et al., 2016; Wang et al., 2016) learn quickly in a limited range of tasks, but that learning is immediately forgotten, such that it must be learned again when the situation arises. The goal of this dissertation is to take steps toward addressing these problems, and the following sections summarize its approach to each problem in turn.

1.1 Incorporating the Environment's Structure

How can we expand the computational theory of RL to account for the inductive biases humans and animals exhibit in natural environments? The recently reinvigorated framework of meta-learning (Thrun and Pratt, 1998; Schmidhuber et al., 1996; Santoro et al., 2016; Vinyals et al., 2016), which focuses heavily on the *task distribution* on which agents are trained, provides an intriguing possibility. In meta-learning, learners are challenged to rapidly learn in an open-ended series of learning problems defined by a distribution over tasks. Critically, it is expected that the learning algorithms the system learns to execute will be shaped at least as much by the nature of the task distribution as they are by the architecture of the learner. In practice, the learners used in meta-learning are generic statistical learning systems, typically deep neural networks. These networks absorb the structure of the task distribution through optimization with respect to that distribution, and are then able to learn

rapidly in new tasks from the same distribution, by making strong inferences based on knowledge of shared structure acquired via meta-learning.

Meta-learning in the past has been used primarily as a tool for building automated systems to learn quickly on new tasks. However, it also provides a process for research in cognitive science. The task distribution can be viewed as an environment model, which should encapsulate the researchers' beliefs about the essential structure of the environment[1]. The architecture should embody a set of algorithmic principles the researcher believes to be important. Finally, the optimization process melds the two, and the resulting learning behavior, including inductive biases, can be compared with those of humans and animals. In essence, this research process allows us to ask whether the behaviors and learning algorithms we see in humans fall naturally out of a small set of architectural components, the characteristics of the environment, and reward optimization. This is the approach followed by the work in this thesis.

This view of meta-learning as a research process is inspired by rational analysis (RA; Anderson, 1990), wherein the researcher first specifies a formal model of the environment, then expresses bounds on the computational capacity of the learner (i.e. limitations of the learning system), then derives analytically the optimal behavior function, which is then compared to human behavior (Oaksford and Chater, 1998)[2]. RA thus explains behavior as resulting from the character of the environment and the bounds on the learner's computational capacity. Meta-learning similarly explains behavior as resulting from the character of the environment and the architectural principles that define the agent's design. In addition to defining bounds versus defining architectural principles, the other major difference is that RA uses analytical methods to find an optimal solution given the environment and bounds, whereas

[1]For past work discussing the statistical structure of the natural environment, see Botvinick et al. (2015). Work on ecological rationality (e.g. Goldstein and Gigerenzer, 2002) is also closely related.

[2]This method also bears resemblance to the loss function and optimization focused approach to neuroscience advocated by Marblestone et al. (2016).

meta-learning uses numerical methods to approximate an optimal solution given the environment and architecture.

In keeping with its inspiration from rational analysis, and its reliance on architectures and optimization methods from the connectionist literature, I will refer to this mode of research that begins with a formal environment model, defines a neural network architecture that realizes algorithmic principles, then uses numerical optimization to derive the (learning) behavior that emerges from the two, from here on as connectionist rational analysis (CRA).

Wang et al. (2016; 2018) applied such an approach - although without describing it as a form of rational analysis - to show that a variety of aspects of learning behavior and neural recordings could be reproduced by generic, model-free recurrent networks trained on specific task distributions (see Section 2.2). This dissertation extends that work, both in terms of the environment model and the agent architecture, in an attempt to explain a superset of the phenomena that theory addresses.

1.1.1 Structure, Novelty, and Reoccurrence

An essential feature shared by all of the environment models used by Wang et al. is constant novelty, generated from persistent structure. Formally, these environments draw Markov Decision Processes (MDPs) m from task distributions $\mathcal{D}$,

$$m \sim \mathcal{D}. \tag{1.1}$$

Wang et al. show that the learning algorithms their agents learn are heavily influenced by the nature of $\mathcal{D}$. For example, in a series of bandits tasks, they show that the agents learn to explore efficiently by making use of inductive biases informed by structure in the bandits problems, such as correlations among the arms. Similar demonstrations are made in the two-step task (Daw et al., 2011), where the agents

learn to execute model-based control through training in an MDP where this is an effective strategy, and in Harlow's learning to learn task (Harlow, 1949), where the agents learn to one-shot learn by leveraging consistent task structure.

This type of environment model recognizes that the world in which animals and humans find themselves routinely poses novel problems, and, that there is shared structure among those problems that enables solutions to be discovered rapidly. However, there is an aspect of natural environments equally pervasive as novelty that this framework misses: reoccurrence. While indeed the natural world regularly poses novel problems, it also frequently presents the same problems repeatedly. This reoccurrence of specific unique challenges may be viewed as a reason that humans and animals need one-shot memory for specific events, i.e., episodic memory. Recognizing that this property of the environment is tightly linked to the episodic learning of interest in this work, one goal of this work was to develop an environment model that captures reoccurrence, in addition to structure and novelty.

To that end, we introduced a framework that generalizes the task distributions of meta-RL using the formalism of stochastic processes (Ritter et al., 2018b). In this class of environment model, tasks are sampled along with identifiable contexts c, which an effective agent can use to identify that a task is reoccurring and retrieve information it discovered previously about the task. These (m, c) pairs are sampled from a stochastic task process Ω with parameters θ and base distribution $\mathcal{D}$, where critically, the Ω is conditioned on the set of tasks that were sampled previously:

$$(m_{n+1}, c_{n+1})|(m_1, c_1), ..., (m_n, c_n) \sim \Omega(\theta, \mathcal{D}). \tag{1.2}$$

This formalism enables the precise definition of the statistics of reoccurrence, enabling the model to incorporate observed statistics of the natural environment (see Section 2.3.2). In summary, we now have a formal model of the environment, in the spirit of Marr's computational level (Marr, 1982), that builds on the abstract

computational framework of RL to model two key features of the natural world: novelty, generated from consistent structure, and reoccurrence. The next section will develop a set of architectural principles suited to environments with these properties.

1.2 An Integrative Account of Episodic Learning

How do humans dramatically change their behavior after a single learning experience, while maintaining the ability to learn slowly and incrementally over time? What mechanisms underlie the striking effects of episodic learning in human behavior, which defy explanation by incremental model-based and model-free theories? How can mechanistic and functional accounts of model-based and model-free learning be integrated with a satisfactory theory of episodic learning? This section combines concepts from episodic control (Gershman and Daw, 2017; Blundell et al., 2016; Pritzel et al., 2017; Lengyel and Dayan, 2008), meta-RL's joint account of model-based and model-free learning (Wang et al., 2018), working memory reinstatement (Hoskin et al., 2018; Cohen and O'Reilly, 1996), and working memory control (Chatham and Badre, 2015b; Kriete et al., 2013; O'Reilly and Frank, 2006; Frank et al., 2001; Braver and Cohen, 2000) to propose answers to these three questions. In particular, the solutions proposed will rely on interactions between the environment and the agent architecture to induce the learning behaviors of interest. See Chapter 2 for further detail.

The algorithmic theory proposed in this dissertation builds on Wang et al.'s (2016; 2018) theory of meta-rl, in which a model-free learning algorithm relying on reward prediction error (RPE) signals carried by striatal dopamine trains a *separate, independent RL algorithm* which operates in the activations of the prefrontal network. The theory is supported by simulations of a variety neuroscientific experiments, wherein the prefrontal network is modeled using an recurrent neural network – specifically a long short-term memory network (LSTM; Hochreiter and Schmidhuber, 1997) –

which is trained to maximize reward using an actor-critic loss function that fills the role of striatal RPEs.

This system is able to learn rapidly in new situations; but, reminiscent of an amnesic patient, as soon as the episode ends, the agent forgets everything that it rapidly learned. Following the observation that the ability to recall such rapidly-learned, one-shot information is attributed to the hippocampal formation, we set out to endow the meta-learning LSTM with architectural features that functionally mimic those attributed to that structure. Specifically, we sought

1. Pattern completion, providing retrieval of information stored in similar contexts (e.g., O'reilly and Rudy, 2001; for review see Rolls, 2013a).

2. Pattern separation, preventing memories from interfering with one another (e.g., O'reilly and McClelland, 1994; for review see Rolls, 2013a).

3. Reinstatement of cortical activity patterns, especially those that support working memory (Marr, 1971; Treves and Rolls, 1994; Cohen and O'Reilly, 1996).

This final requirement is especially apropos: because meta-RL's rapid learning is carried out in the activations of the prefrontal working memory, everything the system learns within an episode is stored in those activations. Accordingly, if those activations can be retrieved and reinstated, the agent will be in a good position to make use of the previously learned information.

Following the CRA approach, we have thus far constructed a formal environment model with essential features of the environment (novelty and reoccurrence), and we have postulated architectural principles we believe to be important. The next step is to create an architecture that realizes these principles which can be optimized with respect the environment model. To do so, we turn to a recent advance in machine learning – the differentiable neural dictionary (DND; Blundell et al., 2016; Pritzel et al., 2017) – which provides the first two requirements. The DND is a key value

dictionary where the lookup keys are embeddings generated by a neural network, and in the case of past work, the values are state-action value estimates. After each timestep, an embedding of the state and the estimated state-action value are appended to the DND's key and value columns respectively. At the beginning of each timestep, the nearest matches to the current state embedding in the DND are retrieved and the associated value estimates are used to select an action.

This system provides two of the three desiderata: pattern completion, by retrieving the values based on pattern similarity in the keys, and pattern separation, trivially by storing each memory to a different slot in the array. However, it does not provide reinstatement of working memory states. To get this feature, we propose to store LSTM cell states in the value column of the DND. When these values are retrieved from the DND, they can be reinstated, provided a suitable mechanism for doing so.

Designing such a reinstatement mechanism must be done with care because, we discovered empirically, naively summing the retrieved activations with the current working memory activations leads to severe problems. This is perhaps unsurprising given prior work demonstrating the necessity to control the flow of information into working memory from perceptual inputs and previous working memory states (Chatham and Badre, 2015b; Kriete et al., 2013; O'Reilly and Frank, 2006; Frank et al., 2001; Braver and Cohen, 2000). Why should information coming from reinstatement be free from this necessity for control? In light of these observations, we found it natural to propose that the same mechanisms used for controlling working memory input and output also control reinstatement. We achieve this reinstatement control in our implementation by extending the usual gating mechanisms of the LSTM to gate one more term: the reinstated activations. To illustrate, consider the usual update equation for the LSTM, which is used to compute the working memory activations (i.e., the "cell" state, in LSTM terminology) at each timestep:

$$\mathbf{c}_t = \mathbf{i}_t \circ \mathbf{c}_{in} + \mathbf{f}_t \circ \mathbf{c}_{t-1}. \tag{1.3}$$

$\mathbf{c}_t$ is the cell state at time t, $\mathbf{c}_{in}$ is a learned function of the perceptual input, $\mathbf{c}_{t-1}$ is the cell state on the previous timestep, and $\circ$ denotes elementwise multiplication. $\mathbf{i}_t$ and $\mathbf{f}_t$ are the "input" and "forget" gates respectively, which are computed as learned functions of perceptual input and previous cell state. To the above update rule we add a new term for the contribution of reinstated working memory states

$$\mathbf{c}_t = \mathbf{i}_t \circ \mathbf{c}_{in} + \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{r}_t \circ \mathbf{c}_{ep} \tag{1.4}$$

Where $\mathbf{c}_{ep}$ is the retrieved state from the episodic memory, and $\mathbf{r}_t$ is the reinstatement-gate, which is computed in the same way as $\mathbf{i}_t$ and $\mathbf{f}_t$, but with a different set of learned parameters. $\mathbf{r}_t$, $\mathbf{i}_t$, and $\mathbf{f}_t$ are thus designed to coordinate among the three contributors to working memory: input from perception, the previous working memory state, and reinstated activations from episodic memory.

So far, I have described an architecture comprised of a recurrent working memory with episodic reinstatement, and a class of formal environment models which use stochastic task processes to define task sequences with structured novelty and reoccurrence. Hereafter I will refer to this architecture and environment model class together as "episodic meta-RL" (EMRL), to highlight their roots in meta-RL (Wang et al., 2018).

1.3 Testing the Theory with the Episodic Two-Step Task

To test whether EMRL can account for the episodic learning behavior humans exhibit, and its arbitration with incremental learning, we simulated an experiment inspired

by Vikbladh et al.'s (2017) episodic two-step task, which enables dissociation of the influence of various learning algorithms on choice behavior (Ritter et al., 2018a). The following paragraphs provide a summary, see Chapter 4 for further detail.

In each trial of this task, participants select between two actions which lead to one of two outcome states; the two actions lead to the opposite outcomes with high probability and to the other outcome with low probability. The outcome states yield reward with probabilities that change over the course of the experiment. Further, trial-unique stimuli are presented at the outcome state of every trial. Previously seen stimuli are shown at the beginning of some trials, and when this occurs, the participant is guaranteed to receive the same reward as on the trial associated with the cue, if the participant enters the same outcome state.

The two stages of each trial make this task amenable to both model-free and model-based strategies. Further, the episodic cues enable learning based on the cue-associated trial (so-called "episodic learning") in addition to learning based on the immediately previous trials (so-called "incremental learning"). An incremental model-free strategy will select the action that was most often rewarded in recent trials. In contrast, an incremental model-based strategy would select the action that was most commonly rewarded *only after common transitions* and was unrewarded after uncommon transitions in recent trials. The episodic counterparts to these strategies operate in the same way, but with respect to the cue-associated trial rather than the immediately preceding trials. Vikbladh et al. (2017) found that participants used a combination of incremental model-free, incremental model-based, and episodic model-based learning.

We found that through training on a task distribution consisting of the episodic two-step task itself, where each sampled task followed a different sequence of randomly walking reward parameters, the EMRL agent learned to execute incremental and episodic model-based RL. Additionally, it was able to arbitrate between the two

learning algorithms, applying episodic learning only on cued trials, and using a combination of episodic and incremental learning on the cued trials.

It is striking that the EMRL agent is able to learn to execute model-based control with respect to the retrieved hidden states' information about the cue-associated trial, because to do so, it is must learn to both extract the outcome state encountered on the cue-associated trial and to select the action most likely to lead to that state. This is a positive indication for future work seeking build architectures that reason over reinstated working memory traces. Another notable aspect of these results is the arbitration behavior, which is accomplished without any explicit architectural mechanism for arbitration. Instead, the parameters of the gates and recurrent connections seem to learn to select between the valuation strategies as a natural consequence of optimization toward tasks that require this sort of arbitration.

Incremental Model-Free Learning and the Formal Environment Model

Another feature of these results worth discussing is the fact that EMRL does not use incremental model-free learning, while humans do. The fact that EMRL learns to use fully model-based learning is consistent with fact that it was optimized for the two-step task itself, in which the model-based strategy garners more reward than the model-free strategy. Human participants in contrast learn in the real world and bring their resulting priors into the lab experiment, where they are tested after minimal exposure to the task. Corroborating this view is the fact that humans become more model-based when trained more extensively on the two-step task (Economides et al., 2015, see Fig. 4b). Additionally, rats which like our agents are heavily trained on the two-step task, also exhibit model-based behavior almost exclusively (Miller et al., 2017, see Figure 2d).

All of this suggests the need for a better environment model (i.e., training distribution) to explain human behavior in this task. Such a distribution should capture

our beliefs about relevant properties; for example, a distribution of tasks (e.g., random MDPs) in which model-free learning often pays off while model-based planning is prohibitively difficult would be a good starting point. A reasonable prediction is that when trained on an appropriate task distribution, the EMRL architecture would learn to exhibit the mixed model-based and model-free strategy found by Daw et al. (2011) and Vikbladh et al. (2017). Future work will be necessary to develop such an environment model and test this prediction by simulation.

Overall, this illustrates one of the strengths of the CRA approach: because the environment model is a first-class component of the theory, it can be updated to reflect empirical data when appropriate. Because behavior is explained as an interplay between architecture and training distribution, effects of prior learning on laboratory experiments fall into the purview of CRA theories.

Section Summary

This section provided, in the form of EMRL, answers to two of the questions raised at its outset: what mechanisms underlie the striking effects of episodic learning in human behavior, and how can algorithmic and functional (arbitration) accounts of model-based and model-free learning be integrated episodic learning. In the next section I'll turn to the third motivating question — how do humans dramatically change their behavior after a single learning experience, while maintaining the ability to learn slowly and incrementally over time — which I'll discuss from the machine learning perspective that drove that thread of work.

1.4 Recall of Rapid Inferences for Deep RL Agents

A second purpose behind the design of the EMRL architecture is to solve a problem of interest in machine learning. In response to limitations presented by the data

inefficiency of early deep learning systems, a popular thread of research has advocated for meta-learning to build knowledge into deep learning systems to support rapid inference (Finn et al., 2017; Wang et al., 2016; Duan et al., 2016; Santoro et al., 2016; Vinyals et al., 2016; Lake et al., 2017). This has made a large impact in both supervised and reinforcement learning settings. However, while these systems are able to leverage knowledge of the general task structure to learn specific tasks rapidly, they are not able to recall what they learn about those specific tasks. As a result, when these systems re-encounter a task, they must solve it again, rather than snapping back to previously discovered solutions. This is true of meta-learning LSTMs (Wang et al., 2016; Duan et al., 2016), memory-augmented neural networks (Santoro et al., 2016), matching networks (Vinyals et al., 2016), and model-agnostic meta-learning (Finn et al., 2017) and its derivatives (e.g., Clavera et al., 2018; Nichol et al., 2018). In Ritter et al. (2018b), we explored the efficacy of EMRL for solving this problem. The following paragraphs summarize the experiments from that paper, and Chapter 3 provides fuller detail.

Contextual Bandits. We first tested EMRL in a contextual extension of the bandit problems from previous work (Wang et al., 2016; Duan et al., 2016). In these contextual bandits problems, the agent is presented with a set of arms, each with a latent reward probability. The agent is given a fixed number of pulls to find and exploit the best arm. During each episode the agent is presented with a stimulus that defines the reward probabilities of each arm. The same stimulus-arm probabilities pair will again be presented in a later episode, at which point the successful agent will recall its past experience with that set of arms, including its inferred solution, and will immediately begin to exploit the best arm. The base distribution of the stochastic task process for this experiment was a uniform distribution over reward parameter-context mappings, which were held fixed during blocks of episodes.

We found that through training on this task distribution, the EMRL agent learned to explore the bandit arms in each context, then recall what it had learned about those arms when it entered the same context later. In other words, after a single exposure to a context, the agent dramatically updated its policy to achieve more reward in subsequent exposures. Further, the agent achieved more reward with each successive exposure to a given context, demonstrating the agent's ability to not only load completed solutions, but also to retrieve partial solutions and pick up learning where they left off. This ability to store action policies then tweak them upon reloading is reminiscent of the algorithm described by (Donoso et al., 2014).

Omniglot Bandits. In the first experiment, the contexts were easily identifiable binary strings. In the second experiment, we added a feature to the task that models an aspect of the real world: when tasks reoccur, they may differ in their perceptual presentation. In this second experiment, reward probabilities were paired with omniglot character classes, so that when a set of bandits reoccurred it would be paired with a different drawing of the same class. This creates an implicit classification problem in retrieving the correct memories. We found that after pretraining the embedding network using an algorithm that operates over contents of a DND, EMRL also solves this challenge.

Compositional Bandits. Next, we asked whether EMRL could handle the more complex, realistic scenario in which tasks don't reoccur wholesale, but instead task features reoccur in new combinations. This models the kind of challenge of interest in prior work in neuroscience (Shohamy and Daw, 2015; Barron et al., 2013). In this experiment, each arm was paired with a stimulus and a reward probability, and arms typically reoccurred in novel combinations. We found that without modifying the architecture, EMRL outperformed the MRL agent by a wide margin in this setting and that it increased its reward with each subsequent exposure to a given context.

Contextual Water Maze Finally, we asked whether EMRL is effective in more sophisticated tasks consisting of multi-state MDPs. To this end, we trained EMRL on a set of water mazes with random starting and goal locations. Stimuli were paired with reward locations and presented along with each maze episode. We found that the meta-RL agents (without episodic reinstatement) performed approximately optimally under the constraint of no knowledge of previous episodes, and that the EMRL agent widely outperformed these agents and this optimum, indicating that EMRL was able to retrieve goal location information from past episodes and execute a navigation policy to that location. Further, the trained EMRL agents navigated to the goal location by one of the shortest paths after a single previous exposure to the context. This last finding is perhaps the clearest example of how a single experience can dramatically change EMRL's policy, while preserving its ability to learn slowly and incrementally when appropriate.

In sum, this thread of work showed that EMRL solves the basic forgetting problem in meta-learning LSTMs, and went further to show that this solution extends to the use of previous task features in a compositional setting. Additionally, it provided evidence that EMRL can explain how humans might update their policies dramatically after a single experience. A number of outstanding issues remain, which EMRL is well positioned to address. First, in this work the query embeddings were not learned online; instead we used pretrained embeddings or trivially identifiable contexts. The differentiable pathway of the DND and contrastive losses over the DND's contents provide promising solutions to this problem. Another issue is that many real-world environments do not provide everything needed to query effectively within one timestep; as a result, it is necessary to aggregate information over time in order to query effectively. To solve this problem, a natural approach is to query using some function of the LSTM cell state. Such an autoassociative retrieval mechanism will also be important for modelling further neuroscientific findings, as discussed in the next

section. Finally, it will be essential to learn mechanisms or find good heuristics for determining when to save, or alternatively, when to forget. The current experimental setup made this problem easy — the agent could be successful by simply saving at the end of each episode. More realistic tasks will not have this property, so it will be necessary to define more robust mechanisms for deciding when to save and/or forget. Future work may pursue (i) heuristics such as saving when reward is received and (ii) learning approaches that leverage curricula beginning with short episodes amenable to backpropagation through the storage process.

1.5 Extensions, Connections and Implications

Having described the model and simulation results, I will now discuss implementational-level extensions of EMRL and EMRL's connections to a range of other findings, theories, and topics.

1.5.1 Episodic Meta-RL in the Brain

The preceding sections develop EMRL as a computational and algorithmic level theory (Marr, 1982). In Chapter 5 (summarized in this section), I endeavor to carry these algorithmic ideas to their logical implementational conclusions given relevant empirical results and previous modeling work. The purpose of the resulting theoretical sketches is to (1) assess the biological feasibility of EMRL's algorithmic processes and (2) provide a framework for relating EMRL for a wider variety of empirical findings and to a broader range of theories and topics in neuroscience. I will first discuss the LSTM and DND separately, and conclude by discussing the coordination between them.

Like Wang et al. (2018) (see also Song et al., 2017) we view EMRL's LSTM as analogous to the prefrontal network, i.e., the PFC and regions in the striatum and

thalamus with which the PFC is strongly connected. The LSTM's primitive computations map onto computations known to be supported by real neurons: linearly weighted summation followed by a non-linear activation function and multiplicative gating (Silver, 2010; Gisiger and Boukadoum, 2011).

The specific set of working memory gates present in the LSTM find analogy in neuroscientific theory as well. Chatham and Badre (2015b) review evidence supporting the notion that cortico–basal-ganglia–thalamic circuits support gating of representations into working memory, out from working memory, and the reallocation of storage capacity in working memory. These correspond to input $\mathbf{i}$, forget $\mathbf{f}$, and output $\mathbf{o}$ in the LSTM. Following this analogy to its conclusion leads to a one-to-one mapping between all of the weights, biases, and activations in the LSTM and synapses and activations in the PFC and basal ganglia. This mapping amounts to a theory that largely agrees with past theories of working memory gating at this level (Frank et al., 2001), with some critical differences (see section 5.1 for details).

Perhaps the most controversial aspect of the LSTM is backpropagation, the learning rule used to optimize it. While there is a considerable literature developing biologically plausible learning rules that use approximate gradient estimates (for review see section 2.2.2 of Marblestone et al., 2016) and some evidence for important roles for retroaxonal signalling (Lewis, 2014; Harris, 2008), the degree to which backpropagation resembles the brain's learning rules remains a matter of heated debate. This thesis remains agnostic to this question, instead providing two ways to view the role of backpropagation in our simulations: (1) The functional effects, in terms of the manner of processing and representation of the trained system, of backpropagation and the brain's learning rules are roughly the same. While this remains a supposition, suggestive evidence for this view comes from the fact that SGD does appear to induce representations similar to those found in the brain (Zipser and Andersen, 1988; Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014). (2) Backpropagation

exclusively serves the purpose of fitting the architecture's parameters to the environment, similarly to the way that an analytical optimization method would be used to derive optimal behavior given a set of bounds and a formal environment model (Oaksford and Chater, 1998). It is only intended to model learning rules at the synapse level insofar as it rationally maximizes reward. Either of these views enable us to learn the lessons models like EMRL have to offer while avoiding unwarranted assumptions about the degree to which backpropagation mechanistically resembles the brain's learning rules (see section 5.1.1 for more detailed discussion).

The DND can be viewed as an easily trainable and scaleable stand-in for an attractor network (Rolls, 2007), and more specifically, a Hebb-Marr-style attractor-based model of the hippocampus (O'Reilly and Munakata, 2000). Like an attractor network, the DND retrieves the stored memory closest to the query under some distance metric Rolls (2010). The Hebb-Marr model stores patterns through Hebbian learning of a recurrent weight matrix, whereas the DND appends patterns to a non-parametric array. The Hebb-Marr model reduces interference among memories sparse projections through the dentate gyrus that precede attractor storage and retrieval in CA3 (a process referred to as *pattern separation*; the DND reduces interference by simply storing each memory in a separate slot. The Hebb-Marr model achieves pattern completion via settling attractor dynamics, whereas the DND achieves pattern completion by computing a distance metric on the query and each stored memory, and returning the nearest one. This analogy between the DND and Hebb-Marr suggests that the DND will be a suitable abstraction for modeling behavior that results from the storage and interference-free, similarity-maximizing retrieval of patterns. In contrast, behavioral variation due to idiosyncrasies of sparsified intermediate representations, Hebbian learning dynamics, or errors in converting pattern-separated representation to their original form will require a more implementationally-detailed model. See section 5.2 for further discussion.

Putting the prefrontal working memory and hippocampal episodic memory systems together requires a mechanism for reinstatement gating in the brain. Fortuitously, empirical work has showcased potential mechanisms for gating of memories retrieved along the hippocampal-PFC pathway. Floresco and Grace (2003) showed that stimulation of the medial dorsal thalamus or the ventral tegmental area modulated the response of prefrontal neurons to stimulation of the hippocampal output tract in rats. This stimulation does not implicate the basal ganglia in exactly the manner expected under a gating theory like that of (Frank et al., 2001), and along with other considerations (see section 5.3), suggests that while the input and reinstatement gating mechanisms may be similar in effect, and may both rely on thalamic modulation, there exist important differences between them. Anderson et al. (2016) hypothesize two other mechanisms for "retrieval suppression" based on connectivity: (1) inhibitory anterior cingulate projections to CA1 via the thalamus might gate retrieval and (2) connections from anterior cingulate to entorhinal gate recall before the query to episodic memory happens. In sum, theory and early empirical data suggest that reinstatement gating is likely to occur in the brain, the exact mechanisms remain an open question and may differ from those underlying the other working memory gating functions (Chatham and Badre, 2015b). See section 5.3 for further discussion.

A more biologically detailed analogue to EMRL that is highly specific at the systems level and moderately detailed at the cellular level may be helpful in guiding experimental work toward answering the question of how retrieval suppression works in the brain. For example, future work could create an intersection of Frank et al.'s model of prefrontal working memory and OReilly and McClelland's Hebb-Marr-style hippocampus model, connecting the two using discussed in the previous paragraph (see section 5.3.1 for details). Such a family of models may be able to generate differential predictions for each gating hypothesis. Section 5.3.1 also sketches another model positioned at a level of detail between that of the LSTM/DND and the Hebb-

Marr and working memory models of Frank et al. (2001) and O'reilly and McClelland (1994) which could be maximally useful for answering another set of important questions.

1.5.2 Empirical Connections

The last several years have seen a surge of interest in the impact of individual episodes on choice behavior, and this has led to a number of empirical findings that an aspiring theory of episodic learning should account for. Chapter 6 discusses the findings of Bornstein et al. (2017); Bornstein and Norman (2017); Wimmer et al. (2014); Duncan and Shohamy (2016); Hoskin et al. (2018) and Einstein et al. (2005), describing their key results, and how EMRL would account for them, or could be modified to do so.

To account for several of these findings, EMRL would need to be modified, or a more biologically detailed instantiation of EMRL will need to be invoked. In making these modification recommendations, I have endeavored to avoid highly specific, bespoke changes for the results at hand; instead the modifications should have general utility for increasing the competence of artificial agents and for modeling human cognition.

The theme that emerges most prominently from these analyses is the need to modify EMRL to query using (some function of) the LSTM hidden states, rather than using a separate perceptual embedding. This change is necessary for EMRL to naturally engage with findings from all six studies, and additionally would bring EMRL closer to the *autoassociative memory* traditionally attributed to the hippocampus (Marr, 1971).

Another theme that cuts across many of these studies is the necessity to move from optimizing the agent for the task itself to optimizing the agent on a model environment that instantiates important properties of the natural environment. Chapter 6 suggests formal environment models (i.e., training distributions) for modeling the

discussed studies, in each case building on an intuition about the environment participants occupy before coming into the lab, and bounded optimal behavior in such an environment produced the inductive biases evident in their behavior in the lab(in an approach taken in previous work by (Plonsky et al., 2015)). An intriguing possibility follows: a single simple training distribution may produce all of the inductive biases observed across a range of studies (see section 6.7 for further discussion), providing a unified normative and algorithmic account for a variety of findings.

A third theme that permeates the analysis of some of these studies is the need for an implementational-level analogue. To address the results of Wimmer et al. (2014), and arguably those of Duncan and Shohamy (2016), it is advantageous to invoke concepts such RPE-modulated Hebbian learning in and storage/retrieval phases in the hippocampus. The DND abstracts over these implementational properties, and so invoking the more detailed models discussed in Chapter 5 is necessary.

In sum, the analysis of EMRL with respect to this batch of relevant empirical studies suggests that (1) EMRL is well positioned to engage with all of the findings if (2) it is modified to query using functions of the working memory state and (3) training distributions based on hypothesized properties of the natural environment are used instead of the task itself. These conclusions point the way for future simulation work with EMRL that may provide a unified explanation for a range of findings.

Predictions

In addition to its potential for explaining previous empirical work, the basic tenets of EMRL lead to a number of predictions which may be tested in future empirical work:

- Pattern reinstatement has been observed during the retrieval of static stimuli (e.g., Bornstein and Norman, 2017); EMRL predicts that such pattern reinstatement should also be measurable during the retrieval of memories that encode sequences of events or the results of computation carried out in WM.

- There is considerable evidence that gating mechanisms regulate the flow of information into and out of working memory (Chatham and Badre, 2015a). Analogous experiments should find evidence for a similar mechanism that gates reinstated activations from long-term memory into working memory.

- It should be possible to modulate the degree of WM pattern reinstatement on a given task by training participants on variants of that task where memory retrieval is either useful or distracting. EMRL predicts specifically that pattern reinstatement in PFC should be affected by such manipulations, while reinstatement in hippocampus should be unaffected.

1.5.3 Relationship with Other Theories and Topics

Because of its stance at the intersection of working memory, episodic memory, and decision making, meta-RL with reinstatement relates to a broad range of other neuroscientific theories and literatures. Chapter 7 discusses implications of a variety of theories and findings for EMRL and vice versa, suggesting along the way future simulation work that leverages EMRL to better understand behavior and/or applies neuroscientific ideas to increase EMRL's capabilities. Discussed topics include free recall and the temporal context model (Howard and Kahana, 2002); working memory and serial recall (Botvinick and Plaut, 2006); rational accounts of memory (Anderson and Schooler, 1991; Todd et al., 2009); gated working memory models (e.g., Frank et al., 2001; O'reilly and Rudy, 2001); planning and preplay (Ólafsdóttir et al., 2018); replay and consolidation (e.g., McClelland et al., 1995); flexible, one-shot generalization (reviewed in Kumaran and McClelland, 2012); novelty detection and the hippocampal-VTA loop (Lisman and Grace, 2005); task sets (Sakai, 2008); and extinction, renewal, and non-parametric state categorization (Redish et al., 2007; Gershman et al., 2010).

1.6 Summary and Roadmap to the Thesis

In sum, this dissertation argues for the sufficiency of a small set of architectural principles along with optimization to maximize reward in an environment characterized by a few specific properties, to explain a wide range of abilities and features of human behavior. The architecture properties are

1. recurrent working memory
2. episodic memory that reinstates working memory activations which are retrieved by pattern completion
3. gating of reinstatement to control influence of episodic memory on working memory,

and the environment consists of a steady stream of tasks which

1. share structure
2. pose novelty and variety within that structure
3. often reoccur exactly.

We find simulation evidence that this framework is able to address the three problems we started with at the beginning of this introduction. First, the CRA approach enables us to build more structure and detail into our computational model of the environment. Second, we find that when optimized to maximize reward in such an environment model, a working memory architecture with episodic reinstatement explains how animals dramatically update their behavior after a single experience, why behavior shows signatures of episodic learning, and how episodic, model-free, and model based learning algorithms may be jointly implemented and arbitrated among. Finally, EMRL provides a promising method for meta-learning methods in machine learning to remember rapidly learned knowledge.

The rest of the dissertation will describe all of this in more detail, beginning in Chapter 2, which explains the formal environment model and agent architecture. Chapter 3 describes a set of simulation experiments that showcase the capabilities of the meta-RL with reinstatement architecture. Chapter 4 describes simulations of the episodic two-step task, which show that the model captures key features of human behavior on the task. Chapter 5 discusses how and where the components of meta-RL with reinstatement might be implemented in the brain. Chapter 6 goes on to discuss how meta-RL with reinstatement might account for a large number of empirical findings, and discusses modifications to the model necessary to do so. Finally, Chapter 7 discusses how the present theory and results relate to a broad range of relevant theories and topics in psychology and neuroscience.

Chapter 2

Episodic Meta-Reinforcement Learning

2.1 Background and Preliminaries

EMRL sets out to explain the functions of and coordination among three different neural systems implicated in reward-driven learning: striatal dopamine, the prefrontal network, and the hippocampus.

2.1.1 Dopamine and Model-Free RL

The first, striatal dopamine, has a long history in the neuroscience of reinforcement learning (RL; reviewed in Glimcher, 2011). A watershed moment in this history came with the publication of work by Schultz et al. (1997) which showed that the firing of midbrain dopamine neurons closely tracks reward prediction error (RPE) signals predicted by temporal difference learning (TD; Sutton and Barto, 1998). In TD learning, estimates of the long-run value of taking a given action in a given state are updated after each state-action-reward trajectory to reflect the reward received. To

illustrate, consider the value update formula for a simple TD learning rule, tabular TD(0):

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma Q(s',a) - Q(s,a)), \tag{2.1}$$

where $Q(s,a)$ is the agent's estimated value of taking action a in state s and s' is the state encountered after taking action a in state s. α is the learning rate and γ is the discount factor. The second term of the formula, $r + \gamma Q(s',a) - Q(s,a)$, i.e., the difference between the anticipated value of $Q(s,a)$ before seeing the reward and after seeing the reward, is the prediction error analogous to that carried by dopaminergic midbrain neurons (Schultz et al., 1997). A large body of work has provided further evidence for the explanatory power of this TD prediction error signal in primates, humans, and rodents (reviewed in Daw and Tobler, 2014).

From the update formula in Equation 2.1, it is clear that TD learning is accomplished without explicit reference to an internal model of the transition structure of the environment. Thus, TD and the striatal dopamine system have been associated *model-free* learning, which traces its roots back to stimulus-response theory (Thorndike, 1911) and relates to habitual action theories developed by e.g. Balleine and Dickinson (1998; for review of these connections see Dolan and Dayan, 2013).

How does the dopamine carrying this prediction error signal cause more adaptive behavior in the future? Dopamine is believed to modulate plasticity at its targets, and a common working theory maintains that a "three-factor rule" is behind dopamine's effect in RL (Wickens 1993; reviewed by Glimcher 2011). The three factor rule holds that synchronous firing of pre- and post-synaptic neurons leads to increased synaptic strength in the presence of dopamine and to decreased synaptic strength in the absence of dopamine. In such a regime, firing that leads to unexpected reward should be reinforced, while firing that led to less reward than expected should be diminished.

2.1.2 Prefrontal Cortex and Model-Based RL

A parallel thread of work has shown that portions of the prefrontal cortex (PFC) encode quantities essential for RL, including action and state values (Plassmann et al., 2007; Padoa-Schioppa and Assad, 2006; Rushworth and Behrens, 2008; Kennerley et al., 2006; Seo and Lee, 2008) and histories of recent rewards and actions (Shima and Tanji, 1998; Seo et al., 2012; Kim and Shadlen, 1999; Matsumoto et al., 2007; Tsutsui et al., 2016; Seo et al., 2007; reviewed by Wang et al., 2018). Characterizing the relationship between these encodings and dopamine RPEs remains an open question, and one tentative answer is that PFC and striatal dopamine subserve two separate RL systems, one model-free, and one model-based (Dolan and Dayan, 2013; Daw et al., 2005; Bromberg-Martin et al., 2010). In contrast to model-free learning, model-based learning involves prospective planning and inference based on the structure of the task. Such model-based algorithms learns by updating estimates of the probabilities of transitioning between states and receiving reward. In the tabular case, this can be accomplished with the following updates:

$$\begin{aligned} T(s_t, a_t, s_{t+1}) &\leftarrow T(s_t, a_t, s_{t+1}) + \alpha(1 - T(s_t, a_t, s_{t+1})) \\ T(s_t, a_t, s') &\leftarrow T(s_t, a_t, s') + \alpha(0 - T(s_t, a_t, s')) \\ R(s) &\leftarrow R(s) + \alpha(r_t - R(s)), \end{aligned} \tag{2.2}$$

where s_{t+1} is the state transitioned to from after state s and s' represents each of the other states in the MDP. T is the learned transition function and R is the learned reward function. Model-based control then selects actions by simulating trajectories of experience and reward using T and R then choosing actions that lead to the most favorable trajectories.

2.1.3 Hippocampus and Episodic Learning

In addition to the PFC and dopamine systems and their associated incremental learning learning algorithms, recent work has demonstrated the influence on human behavior of episodic learning (Vikbladh et al., 2017; Bornstein et al., 2017; Bornstein and Norman, 2017; Wimmer et al., 2014; Wimmer and Buechel, 2016; Duncan and Shohamy, 2016) of the kind associated with the hippocampus (Gershman and Daw, 2017). For example, Bornstein et al. (2017) reanalyzed data from a classic paper by Daw et al. (2006) and found that an episodic sampling model fit the human behavior, RPEs (measured by fMRI), and chosen value (also measured by fMRI) better than the incremental average model used in the original study. Bornstein et al. (2017) then performed a similar experiment augmented with unique images presented along with some trials which were presented again in recall tests interspersed among later trials. They found a significant effect of reward during the first presentation of the image on choice after the recall test. Further, a sampling model that took the cued trial into account fit the choice data better than a standard incremental model. This and other empirical studies along with computational observations (Gershman and Daw, 2017; Lengyel and Dayan, 2008) have motivated increased interest in such episodic sampling models. To illustrate how these work, consider the following basic episodic control algorithm from Blundell et al. (2016) which computes state-action values as

$$Q(s, a) = \sum_i w_i q_i, \tag{2.3}$$

where i indexes state-action value pairs (s, q), each associated with a single past experience, stored in a non-parametric episodic memory. q_i is the state-action value estimate in the i-th (s, q) pair and w_i is a weight on the value estimate q_i that reflects the similarity between the current state s and the state associated with q_i in the memory store. w_i is defined as

$$w_i = k(s, s_i) / \sum_j k(s, s_j), \tag{2.4}$$

where k is a kernel such as cosine or euclidean distance and j indexes the retrieved (s, q) pairs not at index i. Q-values q to be stored in the memory after experiencing a state transition and reward trajectory are defined as

$$Q(s, a) = r + \gamma \max_{a'} Q(s', a') \tag{2.5}$$

where a' are all of the actions available in s', the state transitioned to from s, and $Q(s', a')$ is computed as in Equation 2.3. γ is the discount factor.

2.1.4 Outstanding Questions

We have now covered three algorithms and neural systems which appear to support learning and decision making in humans. This multiplicity of control systems begs two critical questions about how these systems relate to one another mechanistically and functionally. The remainder of this chapter will be aimed at developing a theory that answers these questions.

1. **Mechanistic Coordination**: How does the neural circuitry underlying these systems relate? Are the PFC, striatal dopamine, and hippocampus carrying out totally modular parallel control algorithms, or do these circuits communicate and coordinate to carry out each algorithm. If so what is the nature of this communication and coordination?

2. **Functional Coordination**: If the brain has – at least – three estimates of value at any given time, how does it arbitrate among them? In other words, how does it weight the contributions of these systems (Daw et al., 2005)?[1]

[1] In short, arbitration among learning algorithms can arise as an emergent effect of model-free optimization of a generic learner (such as EMRL) toward tasks that require a combination of valuation

2.2 Meta-Reinforcement Learning

Let's begin with the question of how the dopaminergic and prefrontal systems relate to one another, in other words, how do the RL-relevant quantities decodable from PFC relate to dopamine RPEs? Empirical results showing that dopamine RPEs reflect value inferred from task structure (Bromberg-Martin et al., 2010; Nakahara and Hikosaka, 2012) and model-based valuations (Starkweather et al., 2017; Sadacca et al., 2016; Daw et al., 2011; as cited by Wang et al., 2018) strongly suggest that the dopaminergic learning system interacts with the prefrontal circuitry that likely underlies model-based valuation. These observations render the mechanistic coordination question more precise, because we can now ask — how do the dopamine RPE's come to reflect task structure?

Recent work by Wang et al. (2018) develops a model that answers this and other questions about the mechanistic coordination between these systems. The model consists of an RL agent that slowly learns from experience on a distribution of learning tasks how to execute learning algorithms. These task distributions are defined over Markov decision processes (MDPs) such that at the beginning of each training episode an MDP is sampled from the task distribution as $m \sim \mathcal{D}$ and is used to train the agent. The task distributions are designed such that each sample m requires the agent to rapidly learn by making inferences possible only with knowledge of structure common among the tasks in the support of $\mathcal{D}$. The meta-RL agent consists of a recurrent neural network that receives as input on each time step a state observation and its action and reward from the previous time step. The network outputs an action and state value estimate on each timestep. The most essential features of this architecture are the reward input, which enables to the recurrent network to learn about the results of its actions without waiting for gradient updates, and the recurrent

strategies. Because this arbitration is not built into the architecture explicitly and is evident primarily in the agent's behavior, it will be discussed it Chapter 4, in the context of relevant simulation results.

hidden state, which stores the histories of recent states, actions and rewards, and any important functions thereof (e.g. action value estimates). Because the training regime leads the architecture to learn by reinforcement learning how to reinforcement learn, the architecture and training regime together are called "meta-RL".

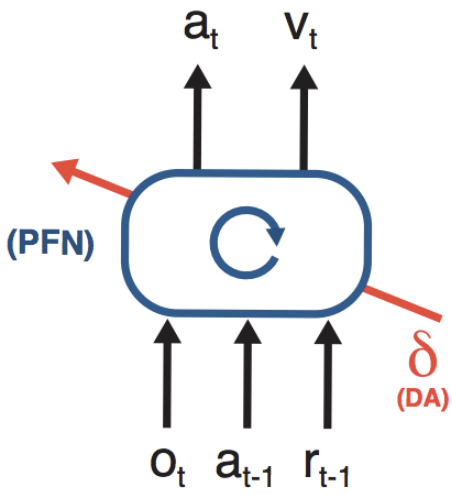


Figure 2.1: Depiction of the meta-RL agent architecture. A recurrent neural network modeling the prefrontal network receives as input a state observation, its previous action, and the reward received on the previous timestep. The network outputs an action and an estimate of the value of the current state. The network is optimized using gradients from an actor-critic loss function which computes a function of the difference between expected and received reward (Equation 2.14) analogous to the reward prediction error associated with dopamine. Reprinted from Wang et al. (2018).

2.2.1 Meta-RL Implementation for Simulations

Wang et al. (2018) simulate meta-RL using a long short-term memory network (LSTM; Hochreiter and Schmidhuber, 1997) as the recurrent working memory. LSTMs are recurrent networks with gates controlling the inputs to, outputs from, and reallocation of the working memory state (for connections to prefrontal gating theories in neuroscience, see Chatham and Badre, 2015b). The working memory, or "cell", state on time step t is computed as

$$\mathbf{c}_t = \mathbf{i}_t \circ \mathbf{c}_{in} + \mathbf{f}_t \circ \mathbf{c}_{t-1} \tag{2.6}$$

where c_{in} is a parameterized function of the sensory input (i.e., state observation) which can be viewed as analogous to representations in sensory cortices. More precisely, c_{in} is computed as

$$\mathbf{c}_{in} = \tanh(\mathbf{W}_{xp}\mathbf{x}_t + \mathbf{W}_{hp}\mathbf{h}_{t-1} + \mathbf{b}_p) \tag{2.7}$$

where $\mathbf{W}$ denotes a weight matrix and $\mathbf{b}$ denotes a vector of biases. $\mathbf{h_{t-1}}$ is the hidden state of the LSTM from the previous timestep. An essential element of the meta-RL setting is that the LSTM input $\mathbf{x_t}$ includes the reward from the previous time step (and optionally, the action from the previous time step). The multiplicative gates $\mathbf{i}_t$ and $\mathbf{f}_t$, defined as follows, coordinate between the contributions of previous working memory state c_{t-1} and incoming perceptual inputs $\mathbf{c_{in}}$.

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i) \tag{2.8}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f). \tag{2.9}$$

The output gate o defined similarly (in Equation 2.10), gates the output from the cell state c_t, as shown in Equation 2.11

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_o + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o). \tag{2.10}$$

$$\mathbf{h}_t = o_t \circ \tanh(c_t). \tag{2.11}$$

In keeping with the meta-RL setup, a policy and value estimate are computed from the cell output using functions (single or multi-layer neural networks) with separate parameters

$$\begin{aligned} \pi_t &= \text{softmax}(f_\pi(h_t)) \\ V_t &= f_V(h_t). \end{aligned} \tag{2.12}$$

Note that the policy π_t is a probability distribution over actions to take. The elements of the vector produced by $f_\pi(h_t)$ are interpreted as logits which are normalized by the softmax function to yield a multinomial distribution over actions. The agent's actions are selected by sampling from this multinomial distribution.

The parameters θ of the LSTM and f_π are trained to maximize reward on the distribution of learning tasks

$$J(\theta) = \mathbb{E}_\pi[r]. \tag{2.13}$$

This optimization is accomplished using the following advantage actor-critic incremental update rule (Wang et al., 2018; see for details and derivation Silver, 2015; Sutton et al., 2000):

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log\pi(a_t|s_t,\theta)[G_N - V(s_t,\theta_v)]. \tag{2.14}$$

α is the learning rate and $\alpha\nabla_\theta \log\pi(a_t|s_t,\theta)$ is the gradient of the log of the policy. G_N is the return over N action-outcome steps in the environment plus the agent's

value estimate of the state encountered at time $t+N$, $V(s_{t+N}, \theta_v)$. $[G_N - V(s_t, \theta_v)]$ is called the advantage function because it represents the advantage the agent obtained by taking action a_t in state s_t, relative its generic estimate of the value of state s_t. (s_t, θ_v) is called the advantage baseline. Note that the actor of this actor-critic algorithm refers to the policy $\pi(a_t|s_t, \theta)$ while the critic component is the value estimation function V. This optimization procedure amounts to training the parameters of the LSTM to maximize reward using an approximate policy gradient loss function and backpropagation (through time)[2].

This actor-critic learning algorithm trains the weights of the network, which is expected to result in emergent learning algorithms carried out by the recurrent dynamics of the LSTM, which use the LSTM recurrent state **c** and **h** to store important inputs and intermediate computed quantities (such as value and volatility estimates).

2.2.2 Meta-RL Simulation Results

Wang and colleagues used this setup to simulate a wide range of experiments. First, they simulated experiments from (Tsutsui et al., 2016) and (Lau and Glimcher, 2005) in which monkeys chose between two actions which held latent reward probabilities that changed throughout the task. Wang and colleagues found that the system closely matched the monkey's behavior and that units in the LSTM coded for the previous action and reward, the conjunction of the two, and choice value. This result provides support for meta-RL's explanation of the relationship between PFC encodings of task relevant variables and dopamine RPEs: the RPEs in a given episode are computed based on value estimates predicated on the variables encoded in PFC, and those encodings are themselves shaped by the RPEs over the course of training during previous episodes.

[2]The loss function includes two other terms, one for training the value estimator $f_V(h_t)$ (gradients from which propagate through h_t to the rest of the LSTM) and for regularizing entropy to encourage exploration behavior. See Wang et al. (2018) Materials and Methods section for the full formula.

Next, Wang and colleagues showed that meta-RL reproduces the structure sensitive RPE signals that challenged the basic view of DA as a purely model-free error signal. This "inferred value" effect was demonstrated in an experiment by Bromberg-Martin et al. (2010) in which a monkey selected between two choices. One of the choices led to reward while the other did not, and this mapping flipped randomly throughout the experiment. The inferred value effect is that when the monkey observed that one of the options leads to an unexpected reward outcome, reflecting the fact that the mapping has flipped, the dopamine response to making the other choice immediately flips as well. Meta-RL reproduces this effect because the reward prediction term in the reward prediction error is determined by the LSTM's value output which, in the trained agent, reflects knowledge of the task structure. Meta-RL thus explains the inferred value effect as arising from learned state representations and state-value estimates in the PFC that feed into the dopamine RPE calculation.

Next, Wang and colleagues showed that meta-RL is able to learn to execute model-based learning through purely model-free RL. To do so they simulated the two-step task of Daw et al. (2011), which affords both model-free and model-based learning (see Chapter 4 for details). In addition to learning model-based control, meta-RL's RPEs reflected the structure of the task, as did the RPEs measured in human ventral striatum by fMRI by Daw et al. (2005).

Having provided explanations for the relationship between dopamine and RL-relevant quantities in the PFC, the relationship between human TD and model-based learning, and for the inferred value effect, Wang and colleagues went on to showcase the degree to which meta-RL's learned learning rules could differ from the generic model-free algorithm in Equation 2.14. To do so they simulated a two-armed wandering bandits task from Behrens et al. (2007) that interspersed periods of volatility (rapidly changing reward parameters) and stability (static reward parameters). Behrens and colleagues found that humans increased their learning rate in the pres-

ence of volatility and decreased it in the presence of stability. Further, they showed that a sector of the PFC encodes volatility estimates. Wang and colleagues showed that meta-RL, after training on a distribution of such episodes, learned to control its learning rate in the same way. Additionally, the LSTM state of the trained agent encoded volatility estimates. Intriguingly, the learning rates the trained model used were orders of magnitude larger than the policy gradient learning rate used to tune the parameters. These results illustrate the ability of meta-RL to absorb structure of the environment to learn to execute a task-tailored RL algorithm that differs dramatically from the generic model-free policy gradient algorithm used to train the system.

2.2.3 Meta-RL Summary

These and a number of other simulation results solidified meta-RL as a strong candidate model of reward-based learning. Meta-RL provides the following answers to the questions about mechanistic coordination raised at the end of the last section:

1. **Q**: Are the PFC and striatal dopamine carrying out totally modular parallel control algorithms? **A**: No, these circuits rely on one another to carry out their roles in RL.

2. **Q**: If the systems coordinate with one another, what is the nature of this coordination? **A**: The PFC encodes the state variables and carries out the computations that underlie sophisticated structure-sensitive, model-based RL. The PFC learns to do this through model-free learning enacted via reward prediction errors, analogous to those signalled by dopamine[3].

3. **Q**: How do the dopamine RPEs come to reflect task structure? **A**: Because the RPEs are predicated on the value estimates computed from state representations in PFC that reflect the structure of the task, the RPE's themselves are also

sensitive to the task structure. This aligns very neatly with previous theory (Nakahara and Hikosaka, 2012).

2.3 Meta-RL with Episodic Recall

2.3.1 Motivation

While meta-RL provided a simple, coherent story about the mechanisms underlying decision making, a number of important questions related to episodic memory and learning remain open. How do humans accomplish very rapid, even one-shot learning that persists over long timescales (Lee et al., 2015)? What neural processes underlie the episodic learning effects observed empirically? The hippocampus appears to play an important role in these abilities and effects Wimmer and Shohamy (2011) — what exactly is the nature of this role? How does the hippocampus interface with the prefrontal and dopaminergic circuits involve in reward-based learning? While there is a vast literature on the hippocampus' role in declarative memory, that research has largely been isolated from work on the neural bases of decision making (see Wimmer and Shohamy, 2011, for exceptions). This has led to a scarcity of theory on the topic (but see Wimmer and Shohamy, 2011). Indeed, meta-RL behaves essentially like an amnesic patient, effectively remembering and manipulating mental representations and making intelligent inferences over short time scales then forgetting what it learned during an episode when that episode ends.

Moreover, the task distributions used in meta-RL, in contrast to the real world, do not require memory for experiences farther in the past than working memory supports. Such task distributions do not provide the kind of environment structure

[3]The most well studied target of dopamine neurons in the context of reward learning and plasticity modulation is the striatum, however, it is well established that dopamine neurons following the mesocortical pathway target the PFC, so it is plausible that dopamine modulates synaptic plasticity in the PFC as well (Wang et al., 2018, supplement).

necessary to learn the sophisticated episodic learning behavior humans exhibit (e.g. Vikbladh et al., 2017), even if the architecture supported this learning.

This thesis sets out to to develop a theory of reward learning that explains a superset of the findings covered by meta-RL by extending meta-RL with formal environment models that encourage the learning of sophisticated episodic learning behavior and with architectural features that support episodic learning. Figure 2.2 provides a summary of this extended form of meta-RL, which is explained in the following sections.

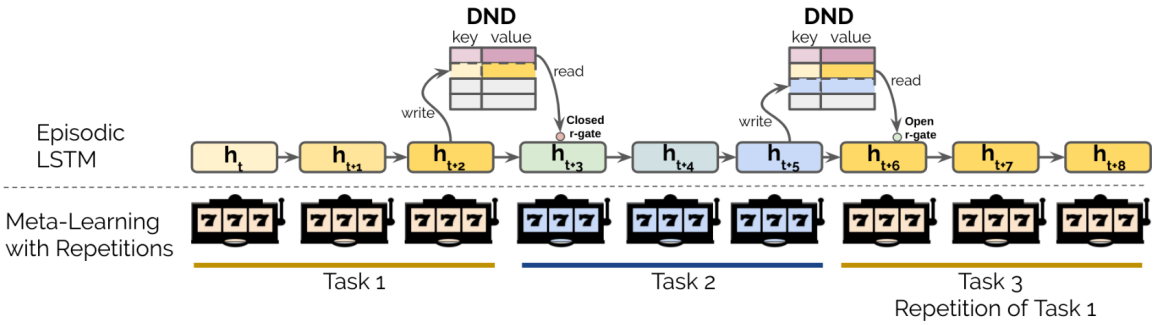


Figure 2.2: Model architecture and environment structure. Tasks, such as multi-armed bandits, are presented sequentially to the agent along with identifiable contexts, which are depicted here as colored bars. On each time step, the agent reads from long-term memory (DND) to retrieve cell states, which are reinstated through the multiplicative reinstatement gate (r-gate). At the end of each task, it writes its cell state to the DND.

2.3.2 Formal Environment Model

I will now describe a formalism for defining sequences of tasks with the kind of reoccurrence that requires episodic memory and learning. Meta-learning research formalized the notion of an unbounded series of tasks by defining task distributions $\mathcal{D}$ over Markov Decision Processes (MDPs), then repeatedly sampling MDPs as $m \sim \mathcal{D}$ Thrun and Pratt (1998). To create open-ended sequences of novel and identifiably reoccurring tasks, we propose to instead sample MDPs m along with contexts c from

stochastic task processes as follows

$$(m_{n+1}, c_{n+1})|(m_1, c_1), ..., (m_n, c_n) \sim \Omega(\theta, \mathcal{D}), \tag{2.15}$$

where Ω is a stochastic process with base distribution $\mathcal{D}$ and parameters θ. n indexes the tasks' positions in the sequence. From here on we will refer to these MDPs and their contexts, (m_n, c_n), as the tasks t_n.

This framework can be used to challenge agents with precisely defined task reoccurrence frequencies. Consider for example a Blackwell-MacQueen urn scheme (Blackwell and MacQueen, 1973), which samples MDP/context pairs t successively as

$$t_{n+1}|t_1, ..., t_n \sim \frac{1}{\alpha + n}(\alpha\mathcal{D} + \sum_{i=1}^{n} \delta_{t_i}), \tag{2.16}$$

where δ_{t_i} is a point mass at t_i and α is the concentration parameter. Intuitively, this scheme first samples an (m, c) task pair from the distribution $\mathcal{D}$ then drops it into an urn. On the following and all subsequent steps, with probability $\frac{\alpha}{\alpha+n}$, the procedure samples a new task from $\mathcal{D}$ and drops it into the urn. Otherwise, it draws a task from the urn uniformly at random, copies it, and drops both the original and new copy into the urn (Teh, 2011). This process has a "rich get richer" property, whereby each occurrence of a task makes it more likely to reoccur, leading to Zipf-like distributions that are observed frequently in naturalistic environments (e.g., Huberman et al., 1998). Task processes like this one may enable the development of agents that can cope with and ultimately take advantage of such important statistical properties of natural environments.

The remainder of this thesis addresses the primary issue of identifying and reusing task solutions when the reoccurrence frequencies are uniform. Accordingly, the bulk of the experiments use the hypergeometric process, which repeatedly samples uniformly

without replacement from a bag of tasks $S = \{t_1 ... t_{|S|}\}$ which contains duplicates of each task, so that $t_n \sim unif(S)$ Terrell (2006). To solve this class of problems, we expect our deep reinforcement learning agents to: (1) meta-learn, capitalizing on shared structure to learn faster with each new task, and (2) avoid exploration when a task reoccurs, instead snapping back to hard-won effective policies.

2.3.3 Architectural Principles and Design

The next challenge is to design an agent architecture that can both explore efficiently based on task structure AND recall what it has learned in a given context when that context reoccurs. To do so, we begin from the observation that meta-RL accomplishes this structure-sensitive exploration fluently and that its episodic learning deficit strikingly resembles that of an amnesic patient. We then set out to design an episodic memory that extends meta-RL, providing some essential functional properties associated with the hippocampus. The hippocampal properties we sought to build into our architecture were

1. Pattern completion, providing retrieval of information stored in similar contexts (e.g., O'reilly and Rudy, 2001; for review see Rolls, 2013a).

2. Pattern separation, preventing memories from interfering with one another (e.g., O'reilly and McClelland, 1994; for review see Rolls, 2013a).

3. Reinstatement of cortical activity patterns, especially those that support working memory (Marr, 1971; Treves and Rolls, 1994; Cohen and O'Reilly, 1996; Xiao et al., 2017).

This final requirement is especially apropos: because meta-RL's rapid learning is carried out in the activations of the prefrontal working memory, everything the system learns within an episode is stored in those activations. Accordingly, if those

activations can be retrieved and reinstated, the agent will be in a good position to make use of the previously learned information.

To implement these features in an efficient, easily trainable system convenient for experimentation, we turned to a recent advance in machine learning — the differentiable neural dictionary (DND; Blundell et al., 2016; Pritzel et al., 2017) — which provides the first two requirements. The DND is a key value dictionary where the lookup keys are embeddings generated by a neural network, and in the case of past work, the values are state-action value estimates (see Figure 2.3). After each timestep, an embedding of the state and the estimated state-action value are appended to the DND's key and value columns respectively. At the beginning of each timestep, the nearest matches to the current state embedding in the DND are retrieved and the associated value estimates are used to select an action.

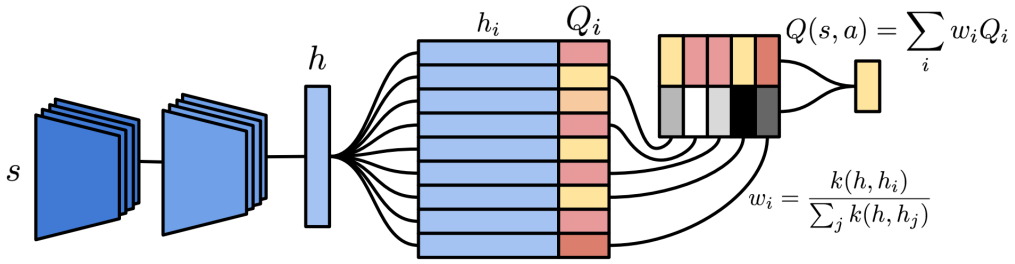


Figure 2.3: The differentiable neural dictionary (DND) with state-action value estimates stored in the values column (Blundell et al., 2016). The state observation s is embedded by a convolutional neural network to produce a query key h. The values paired with the nearest neighbors h_i in the DND are aggregated to produce a value estimate for state s. Reprinted from Blundell et al. (2016).

This system provides two of the three desiderata: pattern completion, by retrieving the values based on pattern similarity in the keys, and pattern separation, trivially by storing each memory to a different slot in the array. However, it does not provide reinstatement of working memory states. To get this feature, we propose to store LSTM cell states in the value column of the DND (Figure 2.4). When these values

are retrieved from the DND, they can be reinstated, provided a suitable mechanism for doing so.

Designing such a reinstatement mechanism must be done with care, because, we discovered empirically, naively summing the retrieved activations with the current working memory activations leads to severe problems. This is perhaps unsurprising, given prior work demonstrating the necessity to control the flow of information into working memory from perceptual inputs and previous working memory states (Chatham and Badre, 2015b; Kriete et al., 2013; O'Reilly and Frank, 2006; Frank et al., 2001; Braver and Cohen, 2000; Badre and Wagner, 2007; Lloyd et al., 2012; Joel et al., 2002). Fortunately, that work also provides a solution to this general problem: gating. Chatham and Badre (2015b) review a long list of empirical and theoretical studies that converge on the notion that the basal ganglia regulates inputs, outputs, and contents of working memory by exerting a gating influence on the pathways into, out of, and the recurrent pathways of the prefrontal cortex. Further, previous work has called attention to the necessity to learn this gating function to control working memory, and suggested that these functions should be learned in a normative manner to maximize reward (especially Todd et al., 2009).

In light of these observations, we found it natural to propose that the same mechanisms used for controlling working memory input and output also control reinstatement. We achieve this reinstatement control in our implementation by extending the usual gating mechanisms of the LSTM to gate one more term - the reinstated activations. This is a natural step, because it makes use of the gating mechanisms in the LSTM that are known to be highly effective and provides the kind of normative gating advocated for by Todd et al. (2009). To accomplish, we add to the above cell state update rule in Equation 2.6 we add a new term for the contribution of reinstated

working memory states

$$\mathbf{c}_t = \mathbf{i}_t \circ \mathbf{c}_{in} + \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{r}_t \circ \mathbf{c}_{ep}$$

where $\mathbf{c}_{ep}$ is the retrieved state from the episodic memory, and $\mathbf{r}_t$ is the reinstatement-gate, which is computed in the same way as $\mathbf{i}_t$ and $\mathbf{f}_t$, but with a different set of learned parameters. $\mathbf{r}_t$, $\mathbf{i}_t$, and $\mathbf{f}_t$ are thus designed to coordinate among the three contributors to working memory: input from perception, the previous working memory state, and reinstated activations from episodic memory.

EMRL expresses answers to the questions raised at the outset of this section.

1. **Q**: How do humans accomplish very rapid, even one-shot learning (Lee et al., 2015) that persists over long timescales? **A**: This can be accomplished by storing working memory states that encode the policy-relevant variables that were inferred and drove the policy during the initial learning experience. By reinstating these activations when a similar context is encountered these variables can immediately drive the policy as they did immediately after they were inferred.

2. **Q**: What algorithms underlie the episodic learning effects observed empirically? **A**: Pattern retrieval mechanisms operating over perceptual state representations retrieve working memory states associated with specific episodes in the past. Task relevant information, such as value estimates, can then be extracted from these episodic traces via the dynamics of and output projections from working memory.

3. **Q**: How does the hippocampus interface with the prefrontal and dopaminergic circuits involved in reward-based learning? **A**: The prefrontal representations that subserve structure-sensitive and model-based RL are stored wholesale in episodic memory. Dopamine trains the gating mechanism that regulates the flow of information from episodic memory to working memory. Note that dopamine

is likely to play at least one other important role in episodic learning, one that becomes important at a higher level of implementational detail than that provided by the EMRL model; see Chapter 5 for further discussion.

The answers provided above are largely silent about how these operations might be implemented in the brain; Chapter 5 discusses an implementational level extension to the EMRL theory that provides suggestions about biological implementation. The next two chapters detail simulations of original tasks and neuroscience experiments from prior work which demonstrate the capabilities and explanatory power of EMRL.

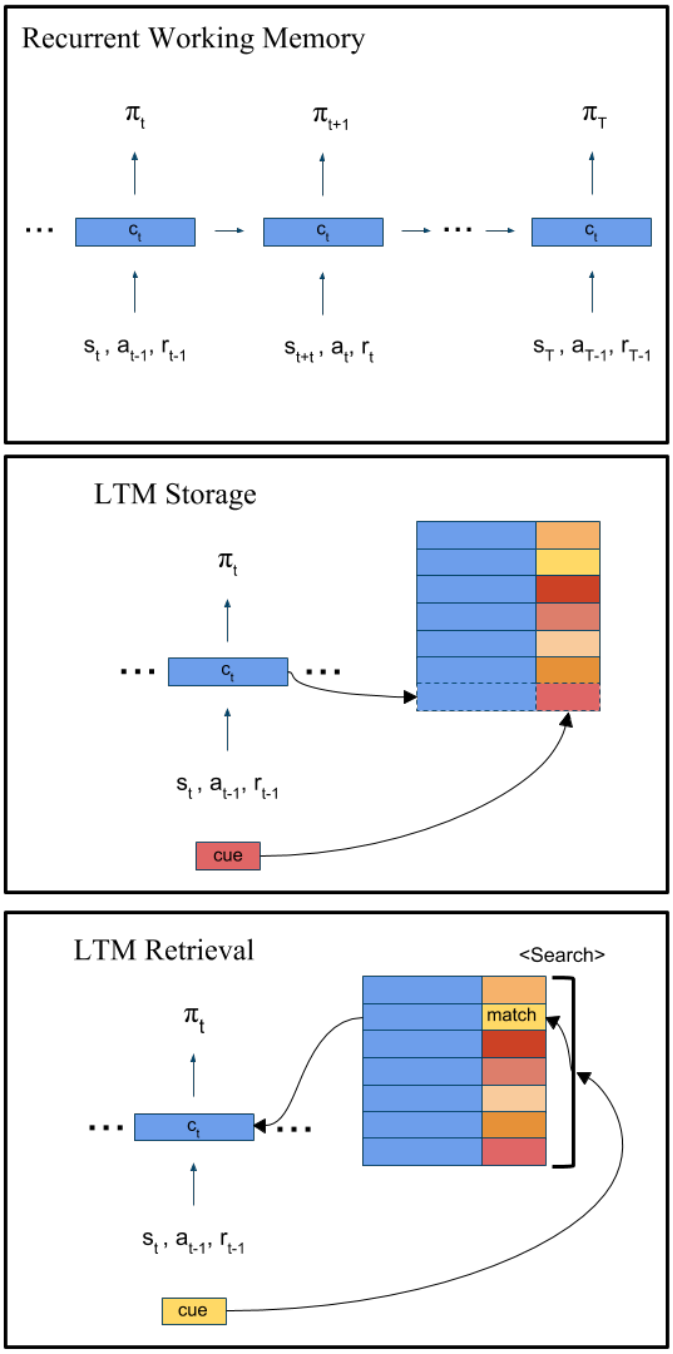


Figure 2.4: (Top panel) The recurrent working memory of EMRL, receiving state observation, previous reward, and previous action as input and outputting a policy on each timestep. (Middle Panel) EMRL's storage process. The cell state of the LSTM is appended to the DND, along with an embedding of the state ("cue"). (Bottom Panel) EMRL's retrieval process. A similarity based lookup between the current state embedding and the stored state embeddings is performed. The cell states associated with the most similar stored state embeddings are retrieved and reinstated to the ongoing working memory process (see Figure 2.5 for details of reinstatement).

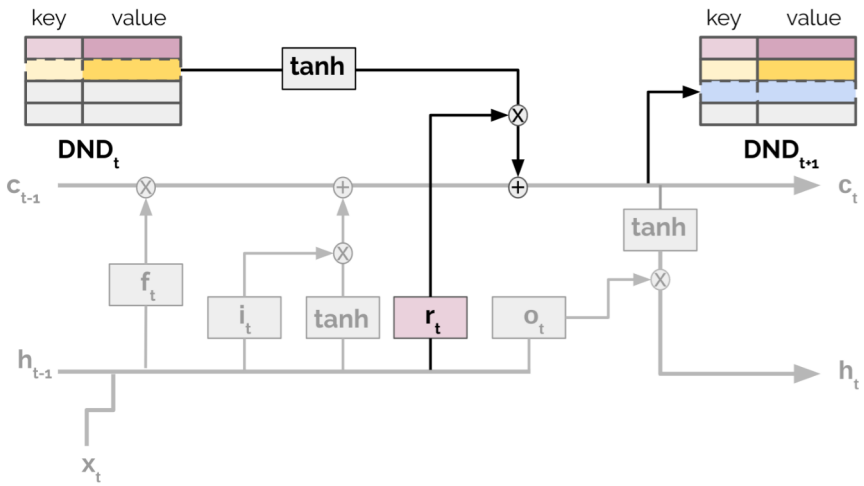


Figure 2.5: A schematic of the LSTM with episodic memory access via reinstatement gate. In gray, the standard LSTM architecture. In bold, the proposed episodic memory and reinstatement pathways.

Chapter 3

The Efficacy of Episodic Meta-RL

This chapter describes 4 experiments which test the capabilities of EMRL. Experiments 1-3 use multi-armed bandits, first exploring the basic case where tasks reoccur in their entirety and are identified by exactly reoccurring contexts (Exp. 1), then, the more difficult challenge wherein contexts are drawn from Omniglot categories and vary in appearance with each reoccurrence (Exp. 2), and then, the more complex scenario where task components reoccur in arbitrary combinations (Exp. 3). Experiment 4 uses a water maze navigation task to assess EMRL's ability to handle multi-state MDPs.

3.1 Experiment 1: Barcode Bandits

Here we address a basic case of episodic repetition, where tasks reoccur in their entirety and are identified by exactly reoccurring contexts. The tasks in this experiment were contextual multi-armed bandits, elaborating the multi-armed bandits of Wang et al. (2016) and Duan et al. (2016). Training consisted of episodes, in each of which the agent faced a set of actions (i.e., arms), where each arm yielded a reward with unknown probability. Each episode consisted of a series of pulls, throughout which the agent should efficiently find the best arm (explore) and pull it as many times as

possible (exploit). During each episode, a context was presented which identified the reward probabilities. The challenge for the agent was to recall its past experience in a given context when the context reoccurred so that it could immediately exploit the best arm – or continue its exploration where it left off.

The contexts in this experiment were binary strings, or "barcodes", so the full set of possible contexts was $C = \{0, 1\}^l$, where $l = 10$ was the barcode length. The reward parameters were structured such that all arms except for one had a reward probability of 0.1, while the remaining arm had a reward probability of 0.9. To prevent agents from overfitting the mapping between contexts and reward probabilities, we periodically shuffled this mapping throughout training. This shuffling serves a similar purpose to the randomization of bandit parameters between episodes: it forces the system to learn how to learn new mappings by disallowing it to overfit to a particular mapping. We hereafter refer to the sequences of episodes in which the mapping between context and bandit parameters is fixed as *epochs*.

We sampled the sequence of tasks for each epoch as follows: we first sampled a set of unique contexts, and paired each element of that set randomly with one of the possible rewarding arm positions b, ensuring that each rewarding arm position was paired with at least one context. We then created a bag S in which each (c, b) pair was duplicated 10 times. Finally, we sampled the task sequence for the epoch by repeatedly sampling uniformly without replacement tasks $t_n = (c_n, b_n) \sim \text{unif}(S)$. There were 100 episodes per epoch and 10 unique contexts per epoch. Thus, each context was presented 10 times per epoch. There were 10 arms, and episodes were 10 trials long.

In this and all following experiments, k was set equal to 1 in the k-nearest neighbors (kNN) search over the DND contents. Cosine distance and normalized inverse kernel were used for the kNN searches (Pritzel et al., 2017). The DND was cleared at the end of each epoch, and the LSTM hidden state was reset at the end of every episode.

Hyperparameters were tuned for the basic MRL model, and held fixed for the other model variations.

Figure 3.1 shows training curves for the barcode bandits task. EMRL widely outperforms the following baselines: MRL, which is the agent from Wang et al. (2016) and MRL+Context, which is that same agent with the barcodes supplied as inputs to the LSTM. The results for this latter baseline indicate that EMRL's performance cannot be matched by parametrically learning a mapping between barcodes and rewarding arms. EMRL's increase in performance over MRL suggests that the memory system is working, and in Experiment 2 we conducted further analysis to verify this hypothesis.

In this and all following experiments, we also tried a variant of the episodic memory architecture in which retrieved values from the DND, c_{ep} were fed to the LSTM as inputs instead of added to the working memory through the r-gate. We found that feeding this 50 dimensional vector to the LSTM reduced performance to random chance. This is almost certainly because the large number of noisy additional inputs make it difficult to learn to use the previous action and previous reward. Because of its poor performance, we omit this variant from the plots[1].

3.2 Experiment 2: Omniglot Bandits

Tasks in the real world are rarely identifiable by exact labels; instead there is often an implicit classification problem to be solved in order to correctly recognize a context. To investigate whether EMRL can handle this challenge, in Experiment 2 we posed an episodic contextual bandits task in which the contexts are Omniglot character classes Lake et al. (2015). Each time the class reoccurs, the image shown to the agent will

[1]We also tried shrinking the dimensionality of c_{ep} using various linear layers and MLPs. We found that only with very aggressive compression (e.g. to <5 dimensions), it was possible to get this to work as well as gated reinstatement on only a subset of our tasks. This degree of compression is unlikely to be suitable for more complex tasks, so we omit further exploration of these architecture variants.

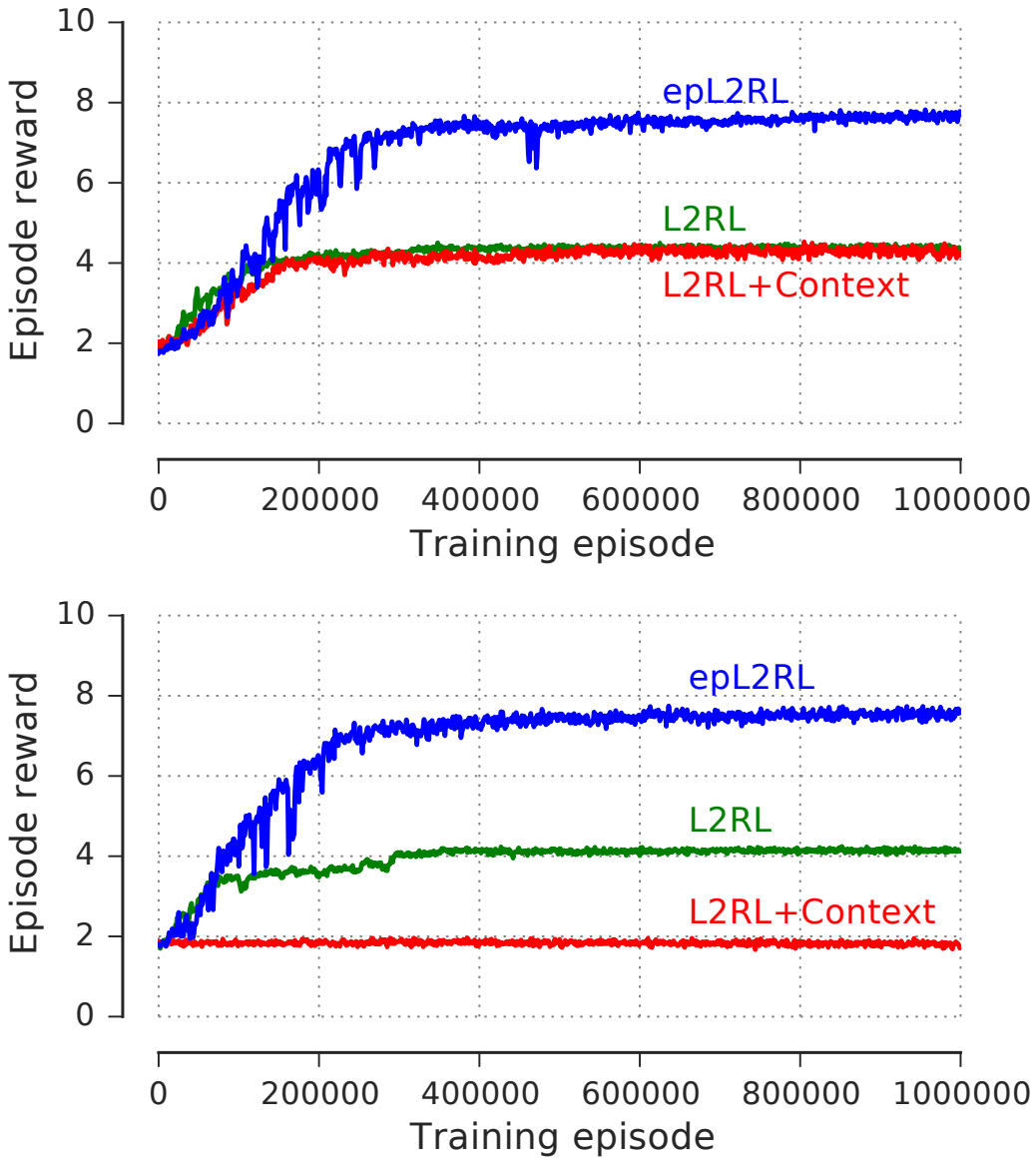


Figure 3.1: (Top) Barcode and (Bottom) Omniglot bandit training curves, averaged over 3 runs.

be a different drawing of the character, so that the agent must successfully search its memory for other instances of the class in order to make use of the results of its past exploration. The task sampling process, episode/epoch structure, and agent settings were the same as those described in Section 3.1.

We used pretrained Omniglot embeddings from Kaiser et al. (2017). This is a particularly appropriate method for pretraining because such a contrastive loss op-

timization procedure Hadsell et al. (2006) could be run online over the DND's contents, assuming some heuristic for determining neighbor status. Future work may try contrastive losses over the DND contents to learn perception online during RL; for discussion see Section 3.5.

The training curves in Figure 3.1 show that EMRL obtains more reward than its MRL counterparts. MRL+Context fails in this case because the Omniglot context embeddings are relatively large – 128 dimensions – drowning out the important reward and action input signals. In Figure 3.2, we analyze the agents' behavior in more detail to understand how EMRL is obtaining more reward. Figure 3.2 (top) depicts EMRL's cumulative regret, i.e. the difference between the expected reward of the optimal arm and the expected reward of the chosen arm, averaged over a set of evaluation episodes. The regret curves are split by the number of previous exposures to the Omniglot character class the agent experienced in that episode. From this we can see that during the first exposure, the agent accrues a relatively large amount of regret. On the second exposure it accrues significantly less, and this trend continues as the number of exposures increases. The possibility that this decrease in regret was due to gradient-based learning is ruled out by the fact that the weights were frozen during the evaluation episodes. This result indicates that the agent is able to recall the results of its previous exploration and to hone them further with each passing episode. In other words EMRL is able to store partially completed solutions then later recall them to "pick up where it left off".

In contrast, Figure 3.2 (bottom) shows that no such improvement with increasing exposures occurs in the MRL agent without episodic memory. Figure 3.2 compares agents' regret curves with those of traditional bandit algorithms: Gittins indices (Gittins, 1979), UCB (Auer et al., 2002) (which comes with theoretical finite-time regret guarantees), and Thompson sampling (Thompson, 1933). We find that the MRL agents compare favorably with these baselines as in Wang et al. (2016) (Figure

3.2 (bottom). Further, we observe that EMRL outperforms these algorithms after a few exposures (Figure 3.2 top), suggesting that it is able to make use of the unfair advantage over these algorithms that its memory affords.

3.3 Experiment 3: Compositional Bandits

Real world agents not only face tasks that reoccur in their entirety, but also task *components* that reoccur in various combinations. This requires that agents compose memories of previously encountered task components. In this experiment we simulate this type of task composition. We continue to use the multi-armed bandit setup, but rather than associating a context with each set of arms, we associate a context with each arm individually. Each episode is then made up of an arbitrary combination of these context/arm pairs. When a context/arm reoccurs, it is not constrained to appear in the same "position" (i.e., it may not be associated with the same action) as when its appeared previously.

This experiment used 2-armed bandits. The sampling process proceeded as follows: at the beginning of each epoch we first sampled a set C_l of unique contexts, and paired each one with the low reward probability of 0.1, yielding pairs (c, l). We then sampled a new set of contexts C_h that contained no overlap with C_l, and paired each member of C_h with the high reward probability of 0.9, yielding pairs (c, h). Next we created a bag S_h containing 5 duplicates of all *high* rewarding contexts/reward probability pairs and a bag S_l containing 5 duplicates of all *low* rewarding contexts/reward probability pairs. Finally, for each episode in the epoch we sampled randomly without replacement a low rewarding context $(c_n, l_n) \sim \text{unif}(S_l)$ and a high rewarding context $(c_n, h_n) \sim \text{unif}(S_h)$. The agent was then trained on an episode with these context/reward probability pairs.

In this setup the stimuli are paired directly with the reward probabilities, and the reward probabilities and stimuli together randomly swap positions on each trial. The stimuli are given to the LSTM as input in positions associated with the action, so that the LSTM must learn to discover the rewards associated with stimuli rather than with the arms. On each trial the EMRL agent queries the DND using one of the two stimuli (selected randomly), retrieves a single cell state from the DND, then reinstates it as in the previous experiments. Otherwise, the agent settings were the same as those described in Section 3.1. There were 1000 episodes per epoch and 400 classes per epoch, so that each class was shown 5 times.

This setup and the aspect of the world that it models pose a specific problem for episodic memory. As in the previous experiments, the agent must retrieve holistic episodic memories based on context similarity; but, in this case the memory state vector the agent retrieves will contain not only information relevant to the task, but also information about another arm which is probably not present. Further, the information the memory contains about the relevant arm may refer to a trial in which that arm was shown in the other position. The agent must extract the relevant information from the memory without allowing the irrelevant information to affect its behavior, and must apply the correct information to the correct arm of the current task. In essence, because the environment state when the memory was saved differs from the current environment state, the agent must map information from the past onto the present.

We find that EMRL performs well even with no architectural modification, as evidenced by its increase in reward over MRL+Context (Figure 3.3 top). Further, the regret curves (Figure 3.3 bottom) show that the EMRL agent consistently reduces its regret after successive exposures to a given context. This decrease in regret cannot be attributed to gradient-based learning because the weights were frozen during the evaluation episodes. MRL (without context) is omitted from Figure 3.3a because in

this task it is impossible to determine the action-reward probabilities without the stimulus input.

3.4 Experiment 4: Contextual Water Maze

In real-world episodic repetition, agents must explore stateful environments while managing contributions from episodic memory. The bandit tasks do not provide this challenge, so we now test our agents in minimal multi-state MDPs, specifically, contextual water mazes. In these tasks, the agent is shown a context barcode and must navigate a grid to find a goal using actions left, right, up, and down. The goal location is initially unknown. When the agent finds the goal, the agent's location is reset (at random), and, if it uses its LSTM working memory effectively, it can proceed directly back to the goal without needing to re-explore. In our setup the grid has 16 states (4x4) and the agent has 20 steps per episode to find the goal and return to it as many times as possible.

The contextual (barcode) cues are mapped to goal locations in the same way they were mapped to rewarding arms in Experiments 1 and 2. Thus, once a goal is reached in a given context, the agent can use its memory of previous exploration to proceed directly to the goal without needing to explore again. The epoch structure and task sequence sampling process were the same as those described in Section 3.1. The agent settings and architecture were also identical, except that (x,y) coordinates were provided as input.

Figure 3.4 (top) shows that the EMRL agent outperforms the MRL baselines. It also significantly outperforms the maximum average reward that could be achieved without memory of past episodes. Figure 3.4 (bottom) shows the mean number of steps for the EMRL agent to reach the goal from episode start, split by the number of previous exposures to the stimulus in the epoch. The means were calculated over

a block of evaluation episodes (with learning rate set to zero). We see that the agent takes a much smaller number of steps to get to the goal after its first exposure. Since the data was recorded while the weights were frozen, this improvement cannot be explained by gradient-based learning, and thus must be the result of information stored in the episodic memory.

Figure 3.5 shows sets of EMRL agent trajectories, again from episode blocks in which weights were frozen. Figure 3.5 (top) shows trajectories associated with the initial exposure to a context in an epoch, and a clear exploration policy is visible. Figure 3.5 (bottom) shows trajectories associated with the second exposure to the same context and starting location. These trajectories show that the agent in all cases navigated directly to the goal by one of the shortest paths.

3.5 Discussion

Over the course of four experiments, we found our agents capable of recalling previously discovered policies, retrieving memories using category examples, handling compositional tasks, and reinstating memories while traversing multi-state MDPs. These results pave the way for future work to tackle additional challenges posed by real-world tasks. First, tasks often will not provide contexts as easily identifiable as barcodes, pretrained context embeddings will not be available, and contexts will be supplied in the same channel as other inputs. As such, a critical next step will be to learn to produce query keys. Two complementary approaches arise naturally for the LSTM with episodic reinstatement. First, the DND provides a mechanism for passing gradients through the retrieval process Pritzel et al. (2017); future work should explore the possibilities of using this learning pathway for EMRL.

Second, the contents of the DND provide an exciting opportunity for auxiliary training procedures, such as Kaiser et al.'s (2017) algorithm. This procedure itera-

tively fills an array with embedding/label pairs. The embedding network is trained by first finding the nearest neighbors in the array to the embedding for a new input image. Then a constrastive loss function Hadsell et al. (2006) is used to minimize the distance in the embedding space between the query and its nearest neighbors that share its label, while maximizing the distance between the query and its nearest neighbors that differ in label. Through this learning process, the nearest neighbors query comes to, with high probability, retrieve only items in the array that have the same label as the query. Because the episodic LSTM's DND already accumulates embeddings and retrieves nearest neighbors, the marginal computational cost of applying such a contrastive loss is relatively low.

Kaiser et al.'s work focussed on supervised learning settings where the "neighborhood function" Hadsell et al. (2006) for the contrastive loss was defined by the true label of the query. In the reinforcement learning the neighborhood function could either be set via heuristics such as temporal contiguity or (whereby states experienced nearby in time are deemed to be in the same neighborhood), more speculatively, it could be meta-learned.

Next, in many tasks of interest, contexts will not be fully observable during a single timestep; instead, information must be aggregated over time to produce an effective query embedding. Consider for example a task in which an agent can identify a previously solved maze only by observing several of its corridors. Using the LSTM cell state, or a function thereof, as the query key is an appealing prospect for handling this challenge. Finally, in the present experiments, episode boundaries were clearly defined so that the agent could simply save at the end of each episode. In the real-world, events are not cut so cleanly, and agents must decide when to save, or save on every step and decide when to forget. Future work may pursue i) heuristics such as saving when reward is received and ii) learning approaches that leverage curricula beginning with short episodes amenable backpropagation through the storage process.

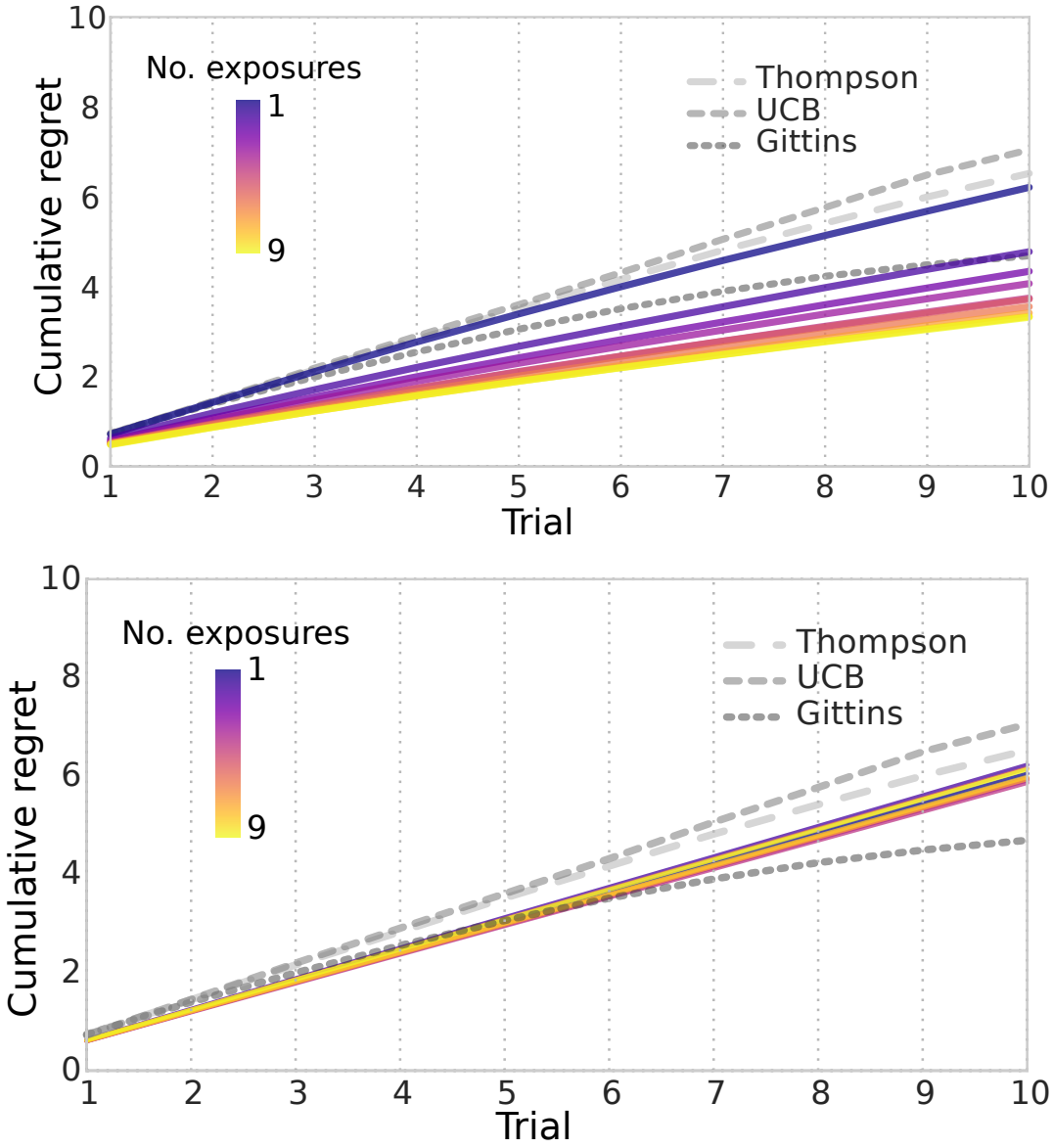


Figure 3.2: (Top) Cumulative regret (expected loss in cumulative reward due to selecting suboptimal arms) for EMRL agent on the Omniglot bandits task, averaged over evaluation episodes. Regret decreases as the number of exposures increases. The agent performs similarly to Gittins indices, Upper Confidence Bounds (UCB), and Thompson sampling on its first exposure, and outperforms them on subsequent exposures. (Bottom) MRL with no episodic memory performs on par with standard algorithms on Omniglot bandits, with no decrease in regret with more exposures.

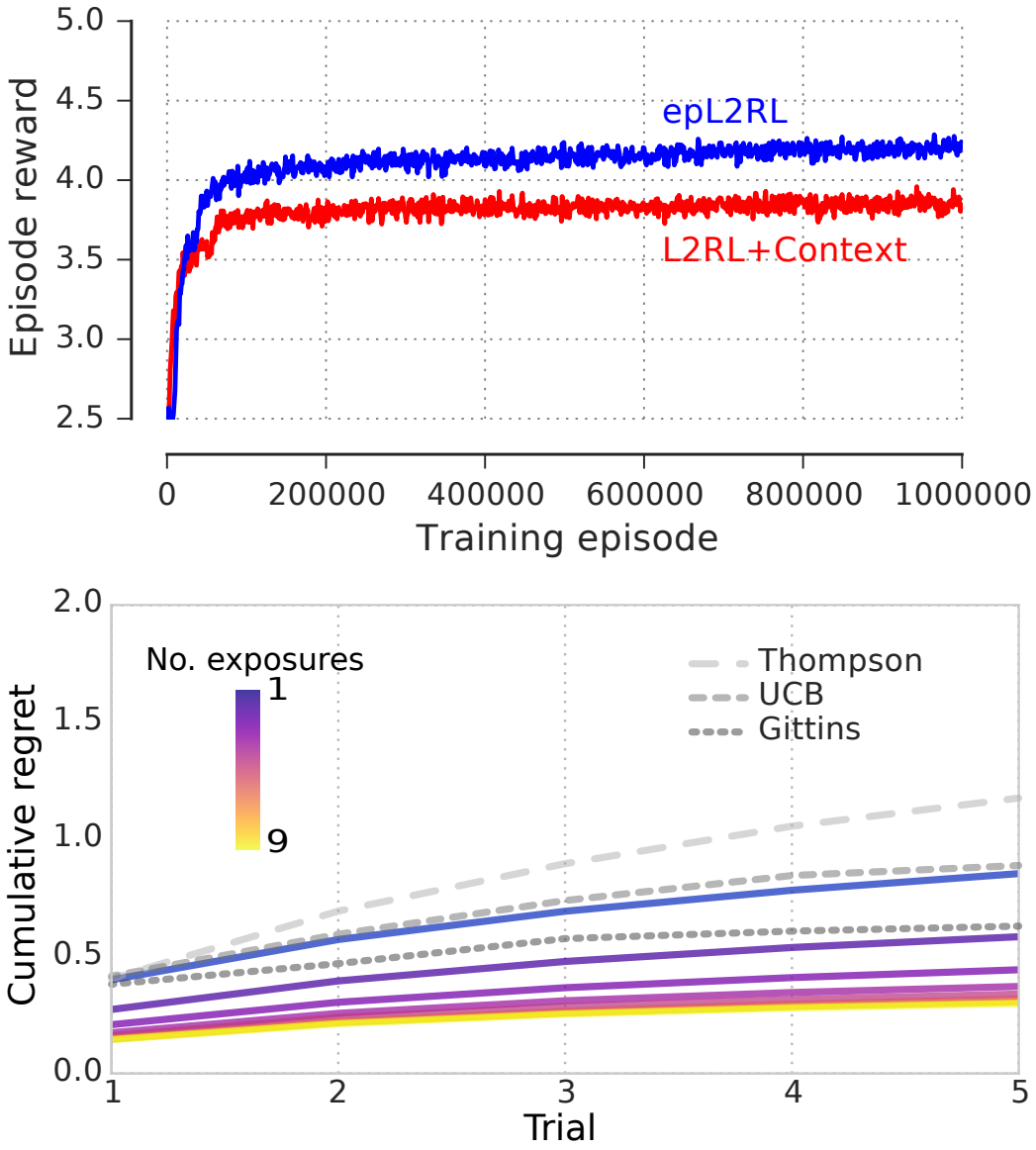


Figure 3.3: (Top) Training and (Bottom) regret curves for compositional bandits. Training curves averaged over 3 runs, regret curves averaged over evaluation episodes.

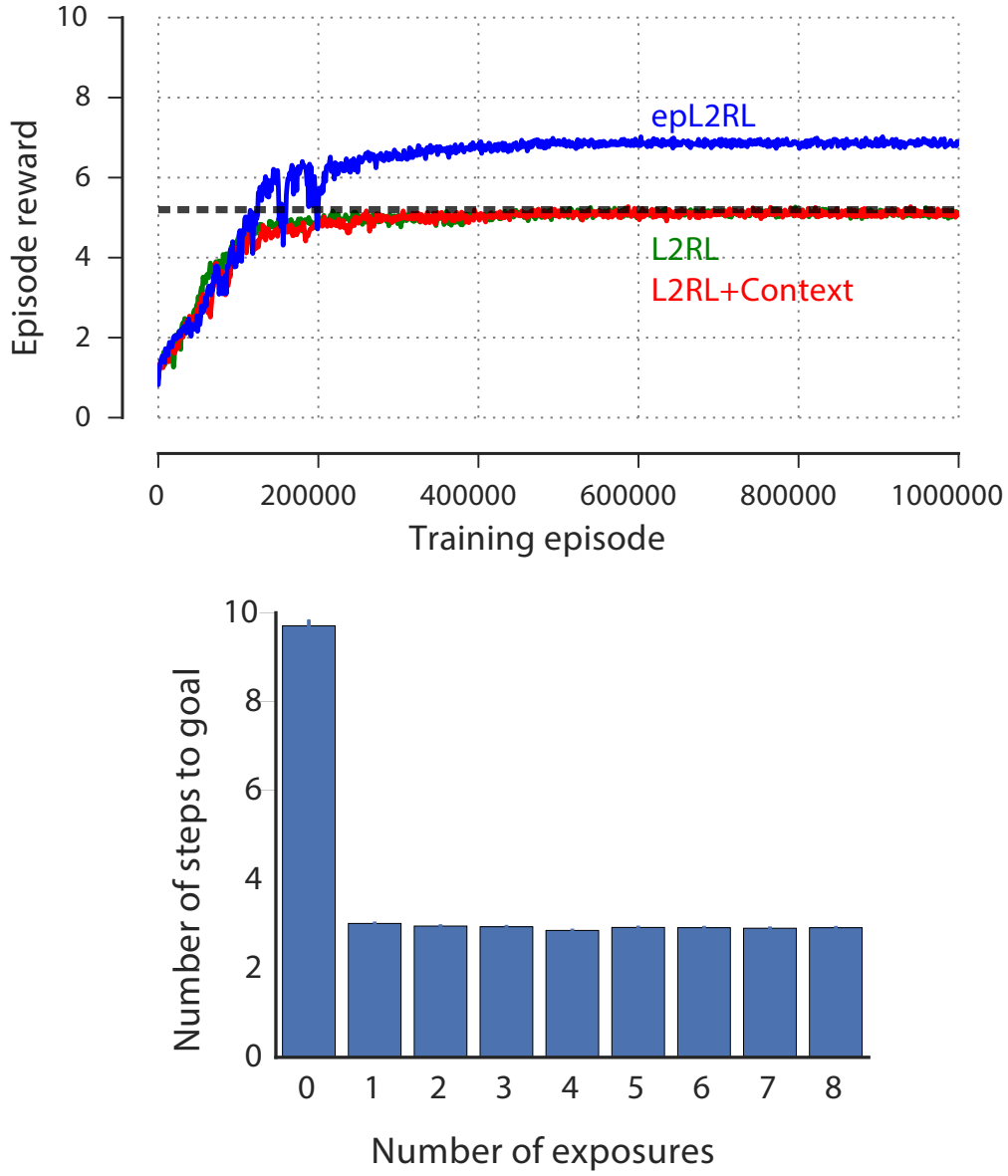


Figure 3.4: (Top) Training curve averaged over 5 runs. The brown line indicates the maximum reward achievable without episodic memory. (Bottom) Number of steps before the EMRL agent reaches the goal after respawning, binned by number of previous exposures to the current context, data from evaluation episodes. After only one exposure, the agent can perform close to optimally.

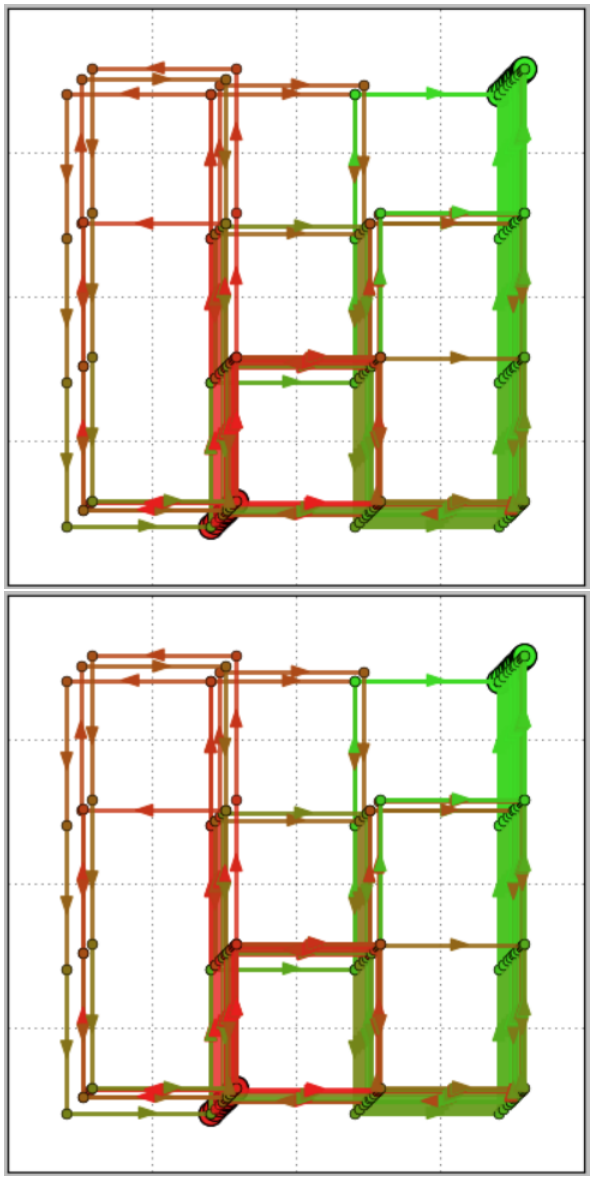


Figure 3.5: (Top) Sample trajectories captured during evaluation for episodes with the same starting position and end goal during the first exposure to a context. (Bottom) Same as c, but for the second exposure. Early steps are colored more red; later steps more green.

Chapter 4

The Episodic Two-Step Task

4.1 Introduction

Nearly every decision an intelligent organism makes is informed by its memory of the results of its past decisions. To be successful, agents must distill the results of past decisions into memories, then make use of those memories to make better decisions in the future. Accordingly, much effort has been directed toward understanding (1) what humans and animals store after a sequence of actions and rewards, and (2) how they use that stored information to appraise the value of future actions.

Model-free and model-based reinforcement learning (RL; Daw et al., 2011; Sutton and Barto, 1998) offer distinct solutions to these two problems. Model-free RL stores statistics about the relationship between states, actions and rewards, and appraises actions by calculating how frequently they led to reward. Meanwhile, model-based RL stores estimated state-state transition probabilities, and appraises actions by using this model to simulate sequences of states to predict future reward. Signatures of both model-free and model-based learning appear in behavior and in the brain (e.g., Daw et al., 2011), and a venerable tradition holds that they are implemented by dissociable neural systems (for review see Dolan and Dayan, 2013).

However, the recent theory of meta-reinforcement learning (meta-RL) proposed that model-free learning, model-based learning, and their sometimes complex interaction could all be explained by a simple unified mechanism (Wang et al., 2018). In meta-RL, a recurrent neural network (RNN) receives a reward signal as part of its input and is trained by model-free learning on a series of interrelated tasks to rapidly learn from this signal. Through this training, the RNN learns to distill the history of observations, actions, and rewards into its hidden state (a form of *working memory*) and to use this summary to select rewarding actions. In essence, this end result constitutes a learned reinforcement learning algorithm that operates in the RNN's activation dynamics. Critically, Wang et al. (2017) showed that this meta-learned RL algorithm can be model-based, even though it was acquired through model-free learning.

While meta-RL provides a full account of incremental learning as it is carried out in working memory, it does not account for the episodic learning processes to which attention has recently been called (Gershman and Daw, 2017). In addition to learning by incrementally storing recent sequences of behavior in working memory, humans appear to learn by storing summaries of individual episodes for long periods of time, then retrieving them when similar contexts are encountered. For example, cues triggering episodic memory retrieval impact reward-based learning, both for good and for ill (Bornstein et al., 2017; Vikbladh et al., 2017; Bornstein and Norman, 2017), and distinctive aspects of episodic memory function contribute to decision-making behavior (Wimmer et al., 2014; Wimmer and Buechel, 2016; Duncan and Shohamy, 2016). Such observations, along with some fundamental computational insights (Lengyel and Dayan, 2007), have recently landed episodic learning a spot beside incremental model-free and model-based reinforcement learning on the list of processes centrally involved in decision making (Gershman and Daw, 2017).

In this present work, we develop a natural extension to meta-RL that enables it to integrate episodic learning. The resulting theory explains how incremental and episodic learning, as well as the coordination between them can be meta-learned through purely model-free RL. The episodic meta-RL theory proposes the following:

1. Meta-RL's working memory is supplemented by an episodic memory which stores working memory states.

2. Each state is paired with a perceptual context embedding that is later used to retrieve the working memory state when similar perceptual contexts are encountered.

3. The retrieved states are then gated into the working memory using a parameterized function, whose parameters are optimized toward the same model-free objective that trains the working memory dynamics.

This proposal is inspired in part by evidence that episodic memory retrieval in humans operates through reinstatement, triggering patterns of neural activity related to those that were induced by the original encoding of the relevant episode (e.g., Xiao et al., 2017), and evidence that reinstatement occurs not only in perceptual systems, but also recreates patterns of activity in neural circuits supporting working memory (Hoskin et al., 2017; Cohen and O'Reilly, 1996). Our implementation of this proposal draws additional inspiration from recent work on differentiable memory systems (e.g., Graves et al., 2016), especially that of Pritzel et al. (2017), which makes use of context-based retrieval for RL.

To empirically test this model, in this work we compared its behavior to that of humans observed by Vikbladh et al. 2017 in a task designed to dissociate the effects of multiple types of incremental and episodic learning. Vikbladh and colleagues found evidence of the use of a model-based form of episodic memory, whereby traces of specific episodes are retrieved from long-term memory based on visual similarity, then

used along with knowledge of the transition structure of the environment to select actions. This episodic model-based learning was present in conjunction with incremental model-free and incremental model-based learning. In the following sections, we describe the task in detail and demonstrate that meta-RL with episodic memory reproduces this model-based episodic learning and its coordination with incremental model-based learning. To conclude, we consider directions for future work, including testable predictions of the theory.

4.2 Task

The task we study is a version of the two-step task (Daw et al., 2011) augmented with episodic cues to previous trials. The task structure, which was inspired by Vikbladh et al. 2017, is diagrammed in Figure 4.1. Each trial consisted of two stages. On the first stage, state s_0, the agent was either presented with the "no-cue" stimulus (a vector of all -1's) on uncued trials, or with a binary vector associated with a previously seen second-stage context on cued trials (see Figure 4.1). In response, the agent chose either a_1 or a_2 and transitioned into one of two second-stage states s_1 or s_2 with probabilities $p(s_1|a_1) = p(s_2|a_2) = 0.9$ (common transition) and $p(s_1|a_2) = p(s_2|a_1) = 0.1$ (uncommon transition). These transition probabilities p were fixed across episodes. On the second stage, the agent was presented with a stimulus representing the context of that second-stage state, followed by a final step in which it was shown the reward outcome. On uncued trials, the states s_1 and s_2 yielded Bernoulli probabilistic rewards of 0 or 1 according to $[r_a, r_b] = [0.9, 0.1]$ or $[0.1, 0.9]$, with the reward contingencies having a 10% chance of switching at the beginning of each trial.

On cued trials, if the agent transitioned into the same state as on the trial k during which the cue was first presented, the agent was given the exact same reward as on trial k. If the agent transitioned into the other state, the reward was determined as

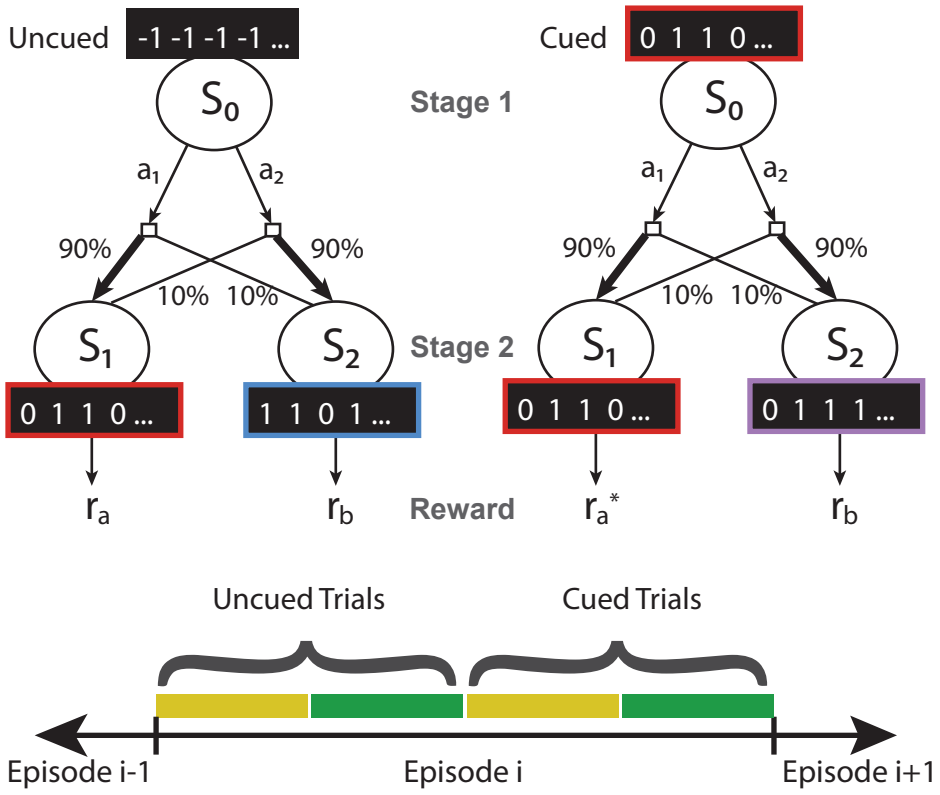


Figure 4.1: Contextual two-step task modeled after Daw et al. (2011) and Vikbladh et al. (2017). (Top) Two trial types are shown: uncued and cued. All trials start in state s_0 at the first stage, at which point agents are presented with either a "no-cue" stimulus or are cued with a second-stage stimulus seen on a previous trial. Transition probabilities after taking actions a_1 or a_2 are depicted in the graph. On uncued trials, s_1 and s_2 result in Bernoulli rewards with probabilities r_a and r_b. On cued trials, transitioning into the same state as in the trial k during which the cue was first presented results in receiving the same reward received on trial k, r_a^*. (Bottom) Trials within an episode are split into 4 blocks, with block 3 consisting of cued trials which are cued with stimuli from block 1, and block 4 cued from block 2.

on uncued trials. This is the critical feature of the task for assessing episodic memory use; if the agent could remember the state entered and reward achieved on the cue-associated trial, it could act to access or avoid that state, depending on whether the past trial was rewarded.

The first half of every episode (50 trials) consisted entirely of uncued trials, and the second half consisted entirely of cued trials, with trials 51-75 being cued with stimuli from trials 1-25 and trials 76-100 cued with stimuli from trials 26-50, randomly

sampled without replacement. This was done to reduce autocorrelation in the reward probabilities by enforcing a minimum of 25 trials between seeing the stimulus on the second stage and being cued with it on the first stage. The agent was trained for 10,000 episodes of 100 trials each, and evaluated with weights fixed on 500 further episodes.

4.2.1 Learning Algorithms

The two-step task with episodic cueing is designed to dissociate among the influences of four different learning strategies on choice. First, the incremental model-free strategy prescribes taking the same action that was taken on the last trial if it was rewarded, and taking the opposite action if it was not rewarded, regardless of whether the transition on the previous trial was common or uncommon. In contrast, the incremental model-based strategy prescribes taking the same action only if the previous trial was rewarded *and* the previous transition was common. If the previous transition was uncommon and the trial was rewarded, the agent will take the opposite action. Episodic model-free and model-based strategies operate like their incremental counterparts, but with respect to the trial associated with the cue rather than the immediately previous trial.

In this task, EMRL's working memory is expected to keep track of the reward probabilities at each outcome state. In order to infer these quantities, it must maintain information about recent actions and states. When the agent receives reward in the final step of a two-step trial, it will save the activations of its current working memory - which encode the agent's outcome state and reward - to its long-term memory. These are saved along with a representation of the context stimulus. At the beginning of a future two-step trial, the agent will encounter this same stimulus. It will then search its long-term memory for matches for that stimulus, and will retrieve the hidden state from the past trial. Crucially, this hidden state will encode the state the agent

encountered at the end of the last exposure to that stimulus as well as the reward received and the action taken. Possessing this critical episodic information, the agent is able to exploit the structure of the episodic two-step task. Specifically, it can learn to implement model-based or model-free valuation with respect to trial information retrieved from long-term memory.

4.3 Methods and Results

We trained EMRL on a task distribution, each sample from which consisted of a block of the episodic two-step task with a random sequence of outcome-state reward probabilities and a randomized episodic cues. After training, we assessed EMRL's performance on a set of evaluation episodes which followed the same structure as the training episodes. To isolate the behavior of the *learned* learning algorithm operating in the activations, all data shown and described in this section were obtained with the weights frozen. To verify that the observed effects are consistent across trained agents, we performed the following analyses on 10 instances of each agent type, where each instance was trained with a different random seed. The reported statistical tests measure the variability across these agent instances. Agent optimization was performed by an implementation of asynchronous actor-critic (Mnih et al., 2016).

First, to determine whether EMRL's episodic memory was performing effectively, we compared the reward EMRL acquired with that acquired by a control meta-RL agent (MRL; Figure 4.2) that was trained and tested in exactly the same way, but did not have access to episodic memory (i.e., its r-gate was always fixed to zero). EMRL achieved more reward overall than MRL ($t(9)=4.28$, $p=1.21e\text{-}4$). This increase in reward came entirely from the cued trials: the difference between reward obtained by EMRL and that obtained by MRL in the cued trials was highly significant ($t(9)=7.205$, $p=1e\text{-}6$), while the difference in uncued trials was not significant

(t(9)=-1.09, p=0.289). These results provide evidence that EMRL was able to use its episodic memory to acquire the additional reward available during cued trials.

Next, we asked whether EMRL exhibited the canonical patterns of incremental model-based and model-free behavior first described in the two-step task by Daw et al. (2011). We formally tested for these patterns of behavior by performing ANOVAs on the probabilities of repeating the previous action, with two binary factors: whether the previous trial was rewarded, and whether the previous trial had a common transition. A main effect of previous trial being rewarded would indicate a model-free strategy, while an interaction between previous trial being rewarded and previous trial being common would indicate a model-based strategy[1]. On uncued trials (Figure 4.4 top), we found a strong effect of the interaction term (F(1,9)=842, p=3.34e-10), indicating that the learned algorithm correctly exploited the transition structure of the task when no episodic information was available. This behavior reproduces the main two-step task result from Wang et al. (2016). The main effect of reward was not significant, indicating the absence of model-free behavior (F(1,9)=3.462, p=0.096). The main effect of transition type was also not significant F(1,9)=2.944, p=0.120). On cued trials (Figure 4.4 middle), we also found an effect of the interaction term (F(1,9)=53.78, p=4.4e-5), indicating that EMRL continued to use the incremental strategy during the cued trials. The main effect of reward was not significant, indicating the absence of model-free behavior (F(1,9)=0.170, p=0.690). The main effect of transition type was also not significant (F(1,9)=5.017, p=0.052).

Next, most centrally, we asked whether EMRL could apply model-based reasoning to information retrieved from episodic memory (Figure 4.4 bottom). We performed the same ANOVA described above, but using as factors: whether the past cue-associated trial was rewarded and whether the past cue-associated trial had a common transition. Since our task guaranteed receiving the same reward if the agent reached the same state as the past trial, the agent should prefer to take the opposite

action as on the past trial if it experienced an uncommon transition and received reward on that trial. We indeed found a strong effect of the interaction term in this analysis ($F(1,9)=36.1$, $p=2e\text{-}4$), and non-significant main effects of reward and transition type ($F(1,9)=4.354$, $p=0.665$; $F(1,9)=4.132$, $p=0.073$). We only performed this analysis on cued trials because the factors are undefined on uncued trials.

To supplement the ANOVA analysis, we fit a probabilistic choice model to EMRL's behavior, similar to the model that Vikbladh et al. (2017) fit to their human data. This model casts participants' choices as a softmax of the weighted sum of valuations from the four valuation strategies. Fitting this model yields estimates of the weight on each strategy (see Supplement for full details). In accordance with the ANOVA results, binomial sign tests on the estimated weights (Figure 4.3) revealed a significant contribution of incremental model-based learning (cued trials, $p=0.002$; uncued trials, $p=0.002$) and episodic model-based learning (cued trials, $p=0.002$), and no significant effect of incremental model-free learning (cued trials, $p=0.754$; uncued trials, $p=0.109$) or episodic model-free learning (cued trials, $p=0.344$).

Overall, the ANOVA and model fit results confirm that EMRL exhibits both incremental and episodic model-based learning, but not episodic model-free learning, in accord with the human behavior observed by Vikbladh et al. (2017). The results show further that, like MRL Wang et al. (2018), EMRL does not use incremental model-free learning, opting instead for the model-based incremental learning that achieves greater reward in the two-step task.

4.3.1 Analysis of Reinstatement Gate Activations

To see what kind of gating function the r-gates learned, we carried out preliminary analyses of the r-gate activations. Figure 4.5 (top) shows timecourses (averaged over 500 episodes) of the r-gate mean for a single trained agent. Each timecourse represents one two-step trial step; that is, whether it was a step where the agent took an action,

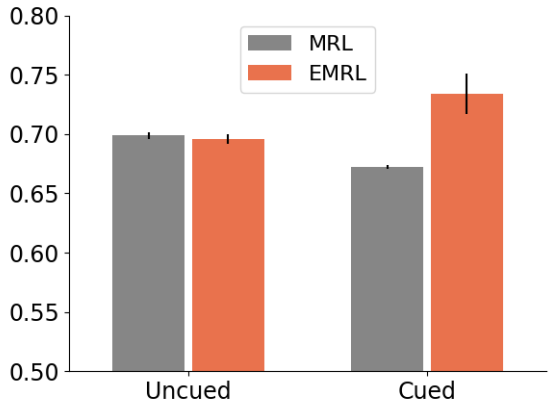


Figure 4.2: EMRL achieves more reward than MRL. Average reward obtained by MRL and EMRL on cued and uncued trials. EMRL earns more reward than MRL on cued trials, suggesting that EMRL can use its episodic memory to exploit the task's episodic structure. For comparison, a random policy achieves 0.5 reward on this task. Bar heights indicate means over 10 trained agents with different random seeds. Error bars indicate the standard error of the mean over the 10 trained agents.

saw an outcome, or received a reward. The timecourses show that the r-gate was more open during cued trials compared to uncued trials, consistent with the presence of useful episodic information on cued trials. Further, the r-gate was most open during the first stage of each trial. This is sensible, because this is the step on which reinstated information can be used to select an action. Figure 4.5 (bottom) compares the r-gate mean for a single trained agent on cued trials when the agent selected the optimal or the opposite action. The r-gate was significantly more open on correct trials, consistent with the idea that retrieved information is necessary to select the correct action on these trials.

While these effects are very clear, and the latter is highly significant, their magnitude is very small. Indeed, the r-gates allow a substantial proportion of the retrieved vector into working memory even during uncued trials when the information is not useful, and this proportion only increases slightly during the cued trials, where the

[1] Akam et al. (2015) demonstrated that there exist sophisticated model-free strategies which can, under the right conditions, produce this reward-by-transition interaction. See Wang et al. (2018), Supp. Figures 6 and 7 for an in depth analysis of this issue with respect to MRL.

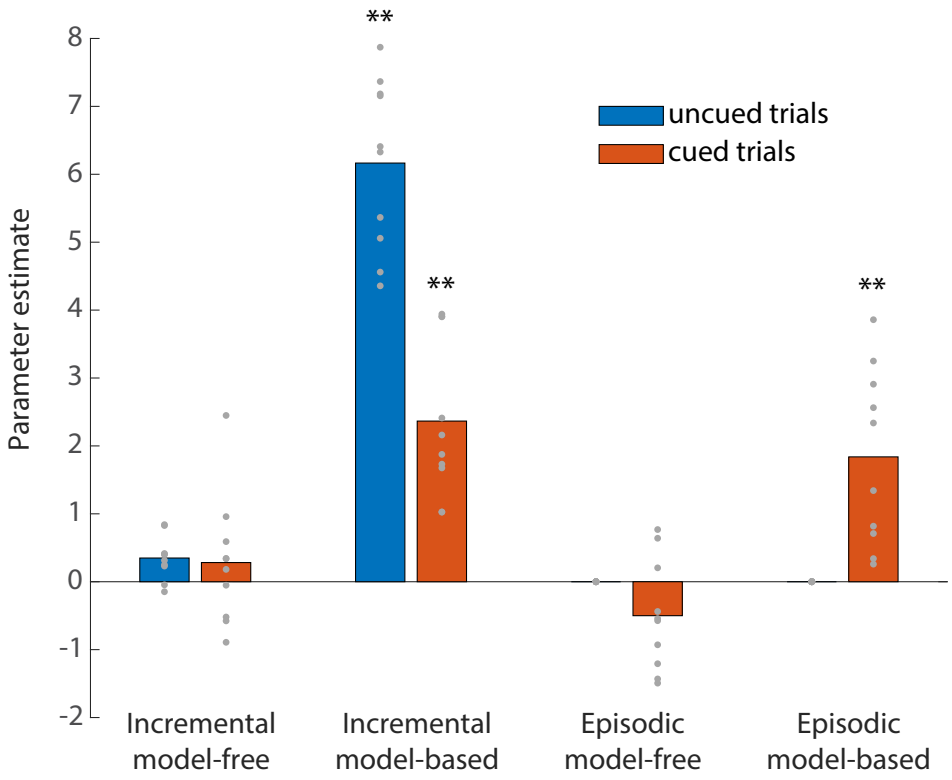


Figure 4.3: Parameter estimates in a model with weighted contributions of four decision systems. The estimates provide evidence that incremental and episodic model-based valuations contribute to EMRL's behavior. The model was fit to EMRL's actions separately on cued and uncued trials. Each gray dot denotes the model fit parameter for one of the 10 trained agents. ** denotes significant difference from 0 at p¡0.005 by binomial sign test.

information is critical. This suggests that mechanisms other than complete gating in/out are at play. For instance, the r-gates may subtly modify retrieved vectors to help the recurrent dynamics and policy layer determine whether or not to use the retrieved information. Further work will be needed to understand the solution the r-gates have found, and the insights gained may extend to biological gating systems. Alternatively, applying regularization, such as dropout, may lead to more easily interpretable neural mechanisms, as in Banino et al. (2018).

4.4 Discussion

The results of this experiment establish that when trained via model-free learning on a task distribution with both incremental and episodic reward structure, EMRL learns to simultaneously execute incremental and episodic model-based learning algorithms. This ability to deploy and coordinate both learning algorithms mirrors that of the participants in the study by Vikbladh and colleagues, providing initial support for EMRL as a model of human decision making. EMRL thus provides an empirically grounded unified account of incremental and episodic learning processes, whereby a single model-free learning mechanism learns to execute and deploy the variety of learning algorithms observed in humans. In addition to support from behavioral data, the model accords in principle with a large neuroscientific literature supporting the notion that episodic memory retrieval recreates patterns of activity in neural circuits supporting working memory (Cohen and O'Reilly, 1996; Staresina et al., 2012; Hoskin et al., 2017; Xiao et al., 2017). Further, EMRL's reinstatement gating mechanism fits naturally with well-supported theories of input, output, and reallocation gating in prefrontal cortex (e.g., O'Reilly and Frank, 2006; Chatham and Badre, 2015a).

It is worth pointing out *why* EMRL learns to use the valuation strategies that it does, and does not learn to use the others (summarized in Figure 4.3). In our view, this follows from the fact that optimized neural systems like MRL and EMRL approach optimal behavior on their training distribution, subject to limitations of the architecture design and efficacy of stochastic gradient learning. In these experiments, the training distribution was the episodic two-step task itself, which yields greater reward to model-based than model-free strategies. EMRL's architecture, with essential features such as associative retrieval and differentiable working memory, enables a broad spectrum of strategies – including model-based and model-free – to be learned via gradient descent. Accordingly, EMRL learns to use model-based instead of model-free strategies because (1) the architecture supports both strategies

and (2) the model-based strategies earn more reward on the training distribution. Future work may investigate ecologically inspired training distributions (Anderson, 1990) that lead generic learning algorithms like MRL/EMRL to reproduce human-like deviations from optimality in laboratory tasks.

In summary, EMRL constitutes a new theory that explains the collage of learning processes observed in humans during decision making as an interplay between working and episodic memory that is itself learned through training to maximize reward on a distribution of learning tasks. Future work may test the predictions made by this model and test the model's ability to replicate additional sources of empirical data.

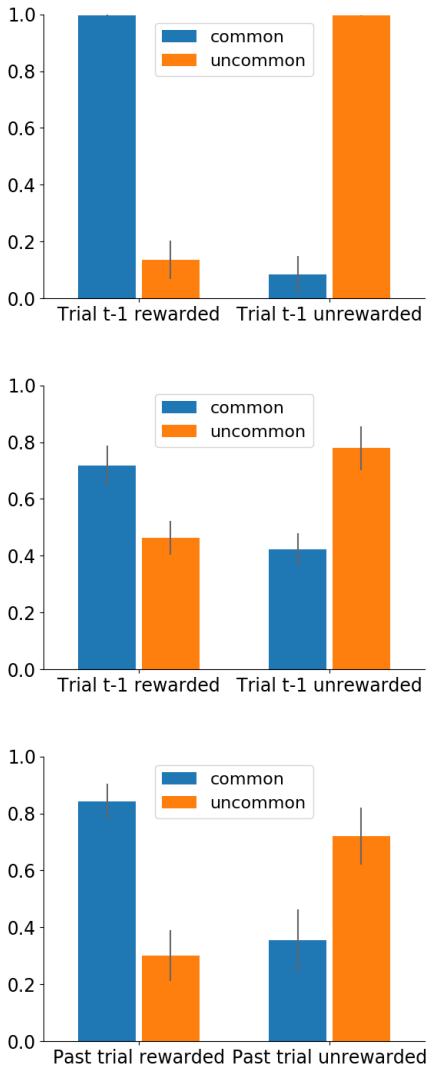


Figure 4.4: EMRL exhibits both incremental and episodic model-based behavior. Bar heights indicate means over 10 trained agents with different random seeds. Error bars indicate the standard error of the mean over the 10 trained agents. (Top) Proportion of uncued trials in which EMRL repeated the action it took on the previous trial $(t-1)$, split by whether it received reward on $t-1$ and whether the transition on $t-1$ was common. The interaction between those two factors is a sign of model-based learning[1]. (Middle) Same as top, but for cued trials. (Bottom) Same as top, but split by whether EMRL received reward on the past trial k when the cue was first encountered and whether the transition on trial k was common.

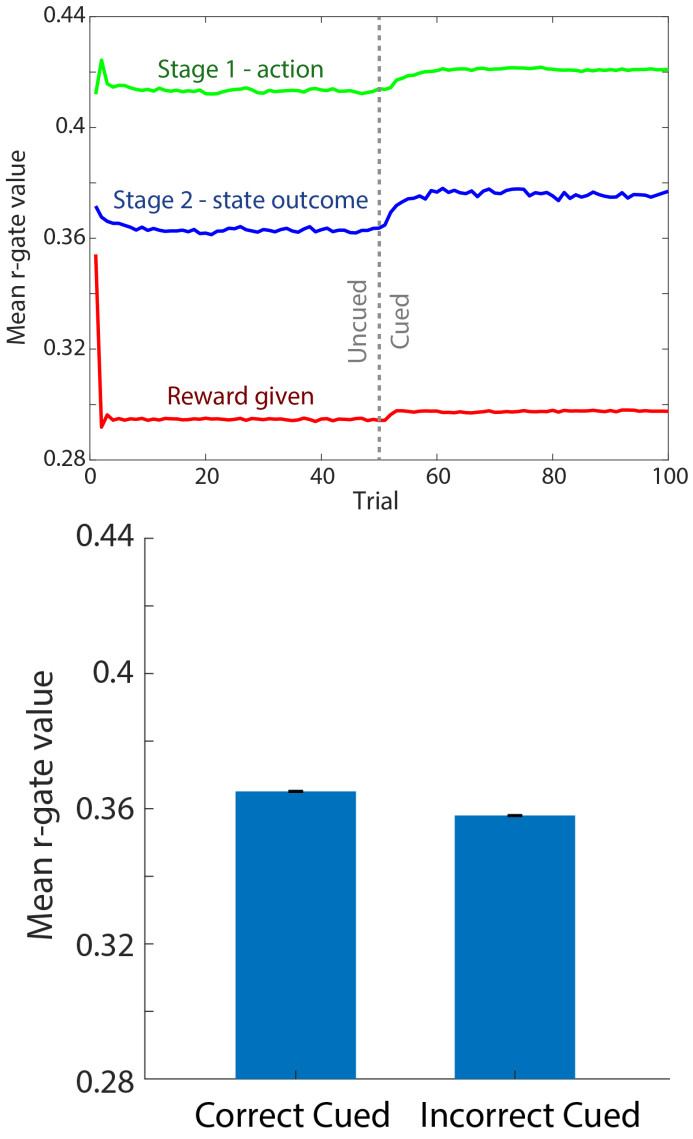


Figure 4.5: (a) Timecourse of the mean values of the reinstatement gate averaged over 500 episodes, split by stage of the trial. (b) Mean values of the reinstatement gate on cued trials on which the agent selected the optimal action (correct) and on cued trials on which it selected the opposite action (incorrect), averaged over all units.

Chapter 5

Episodic Meta-RL in the Brain

The preceding chapters position EMRL as a computational level theory, where task distributions function as formal environment models to which agents are rationally adapted; and as an algorithmic level theory, where concepts such as pattern completion, pattern separation, and reinstatement are abstracted from their biological implementation and designed into experimentally convenient architectural components like the DND. In this chapter I discuss how these algorithmic concepts might be implemented in the brain, providing analogies between each component of the architecture used in simulations and processes that plausibly occur in specific anatomical regions. The purpose of this discussion is to provide a framework for relating EMRL to a wider variety of findings (see Chapter 6) and to other theories and topics (see Chapter 7). This chapter first discusses biologically more realistic analogues to each component of EMRL in turn, then describes how these more realistic components might fit together into an implementational-level counterpart to EMRL.

5.1 Prefrontal Working Memory and the LSTM

Following Wang et al. (2018), we take EMRL's LSTM to represent the following anatomical regions: the prefrontal network (PFN), which includes the prefrontal cor-

tex (PFC; including orbitofrontal, ventromedial frontal, cingulate, dorsolateral prefrontal and frontopolar regions) and subcortical structures to which the PFC is heavily connected including ventral striatum, dorsomedial striatum, and mediodorsal thalamus (Wang et al., 2018). This modeling approach follows a nascent tradition of using homogeneous recurrent neural networks as models of this set of regions, abstracting away much of their physiological and functional differences (Song et al., 2017).

The LSTM uses a small set of computational mechanisms, each of which bears some analogy to known biological neural function in these anatomical regions. Perhaps the most primary mechanism in the LSTM is the linearly weighted summation followed by a non-linear activation function it uses to compute the cell state and gate values (see Chapter 2 for full formulas and description). This mechanic is shared by all connectionist learning systems, and emulates the convergent synaptic summation and firing threshold observed in real neurons, albeit abstracting away details about spike timing.

It is also worth mentioning that the discretized, activation-based LSTM makes it straightforward to maintain activity over multiple timesteps — any recurrent weight matrix with non-zero values on the diagonal (along with non-zero forget gate values) will accomplish this. However, in continuous time, spike based neural networks, this activation maintenance is less straightforward. Fortunately, this has been an exceedingly active topic in systems and cellular neuroscience, so there is a wealth of empirical and theoretical elucidating how this may be accomplished (for review see Zylberberg and Strowbridge, 2017).

The other essential ingredient is the multiplicative gating operation. It is well established that multiplicative connections exist in sensory systems and support functions such as gain modulation (Silver, 2010), and more recent evidence has supported the notion that multiplicative gating regulates the passage of action potentials in the PFC, basal ganglia, and thalamus (Gisiger and Boukadoum, 2011).

At the anatomical level, there exists a very clean analogy between the LSTM's set of gates and the set of gating functions theorized to support working memory. At the heart of this gating theory (Chatham and Badre, 2015b) is the basal ganglia. Originally established as a disinhibitor of motor programs (Mink, 1996), more recently the basal ganglia has been implicated in gating cognitive representations (Frank et al., 2001; O'Reilly and Frank, 2006; Chatham and Badre, 2015b; Hazy et al., 2007). The theorized mechanism by which the basal ganglia gates motor programs is as follows: motor representations in cortex trigger the firing of striatal neurons which project inhibitory connections to the substantia nigra and globus pallidus which in turn project inhibitory connections to the thalamus. Neurons in the substantia nigra and globus pallidus are tonically active, so that the default is for the thalamus to be inhibited. The firing of striatal neurons has the effect, via inhibitory connections, of "releasing the brakes" on the thalamus. In this event, excitatory thalamocortical projections increase activity in cortex, leading to the motor program's execution (Chevalier and Deniau, 1990, as reviewed by Frank et al., 2001). Figure 5.1 shows a diagram of this process.

Because this motor gating function ultimately encourages or discourages activations in motor cortex, the possibility that this same mechanism could apply in prefrontal cortex, and its associated cognitive working memory representations, was quickly identified (reviewed in Frank et al., 2001). Chatham and Badre (2015b) review evidence supporting the notion that such cortico–basal-ganglia–thalamic circuits support gating of representations into working memory, out from working memory, and the reallocation of storage capacity in working memory. Evidence for input gating comes from a variety of human neuroscience experiments. Chatham and Badre present suggestive results of a meta-analysis which showed that studies whose abstracts contained the word "updating" were significantly more likely than those without to report activation in the basal ganglia. Further, training participants in an

updating task increased performance on other tasks that induce striatal response as measured by fMRI (Dahlin et al., 2008). In another study, this type of training led to measurable changes in the timecourse of striatal activation after an updating challenge. Also, (Frank and O'reilly, 2006) showed that pharmacologically silencing NoGo neurons in the striatum enhanced participants' updating ability.

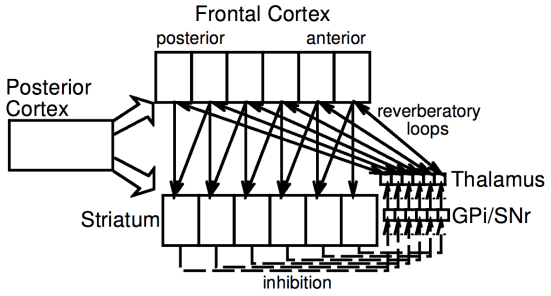


Figure 5.1: Diagram of the basal-ganglia–thalamic–cortical gating mechanism theorized to support working memory updating. Reprinted from Frank et al. (2001).

Support for Chatham and Badre's output gating hypothesis comes from computational models that extend the rostral to caudal frontrostriatal loops that provide motor gating to working memory output (Chatham and Badre, 2015b). The only direct empirical work comes from Chatham et al. (2014), wherein participants undertook a task that posed both working memory updating and output gating demands. Chatham and colleagues found that on trials that required output gating, functional coupling between an area of frontal cortex active during the task and basal ganglia increased.

Empirical work investigating the gating of recurrent connections suitable for reallocating working memory space is relatively minimal. Chatham and Badre (2015b) make the case that reallocation is essential, to do so it is necessary to keep track of the utility of items stored in working memory, and that the basal ganglia is a sensible candidate *a priori* because of its established role in value coding. The empirical evidence for this position consists of a study by Chatham and Badre (2013) which analyzed trials from Chatham et al. (2014) wherein the context was presented in the

middle. On these trials, when the context indicated that the preceding stimulus was irrelevant, allowing more time between the middle stimulus and the last stimulus provided significant gains in performance. This result was interpreted as indicating that subjects required time to reallocate the capacity taken up by the irrelevant stimulus. Chatham and Badre (2013) then provide evidence from an fMRI experiment that the striatum tracks the value of the irrelevant items.

In sum, evidence suggests that the basal ganglia exerts a gating influence on projections to, from, and reallocation of working memory. This trio of gating functions maps cleanly onto the gates of the LSTM, the input, output, and forget gates. Figure 5.2 provides a simplified diagram of the basal-ganglia/prefrontal gating system useful for highlighting this mapping. The arrows from the basal ganglia (BG) to PFC are those discussed in the preceding paragraphs, while those projecting from PFC and from perceptual input to the BG are well established anatomical connections (Purves et al., 2001, Figure 18.4). Recall the LSTM equations from Chapter 2,

$$\mathbf{c}_t = \mathbf{i}_t \circ \mathbf{c}_{in} + \mathbf{f}_t \circ \mathbf{c}_{t-1} \tag{5.1}$$

$$\mathbf{c}_{in} = \tanh(\mathbf{W}_{xp}\mathbf{x}_t + \mathbf{W}_{hp}\mathbf{h}_{t-1} + \mathbf{b}_p) \tag{5.2}$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i) \tag{5.3}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f) \tag{5.4}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_o + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o) \tag{5.5}$$

$$\mathbf{h}_t = o_t \circ \tanh(c_t). \tag{5.6}$$

The cell state c can be viewed as activations in the PFC. c_{in} can be viewed as a representation of the environment state represented in activations in sensory cortices, represented in Figure 5.2 by the symbol s for state. The gate activations $\mathbf{i}_t$, $\mathbf{f}_t$, and $\mathbf{o}_t$ can be viewed loosely as spike rates of cells whose bodies are in the basal ganglia. The

axonal synapses of these neurons would then correspond to the orange bars gating the s to PFC connections, the PFC's recurrent connections, and the PFC to output π connections. $\mathbf{h}_t$ corresponds to the output π, which represents projections out of PFC that influence behavior, for instance, projections to motor cortex.

Following this set of analogies, the pointwise multiplication $\mathbf{i}_t \circ \mathbf{c}_{in}$ could be effected by multiplicative synapses (Silver, 2010) in PFC[1]. This contrasts with theories of gating-by-inhibition in the basal–ganglia-thalamic loop developed by e.g. Frank et al. (2001). This thesis remains agnostic to which of these mechanisms (or a combination thereof) underlies the basal ganglia's gating function. If multiplicative connections hold, then the LSTM can be taken as a somewhat more detailed model; if inhibition via the thalamus holds exclusively, the LSTM should instead be taken as a somewhat more abstract model. Similarly, pointwise multiplication $\mathbf{f}_t \circ \mathbf{c}_{t-1}$ could be effected by multiplicative synapses on recurrent connections in PFC, output gating $o_t \circ \tanh(c_t)$ could correspond to multiplicative connections either in PFC or motor cortex. The alternative hypothesis that other regions (such as the thalamus) and inhibitory connections may be at work instead of direct multiplicative connections remains open for these gating functions as well.

The remaining components of the LSTM to be mapped to anatomical and neural processes are the weights and biases. $\mathbf{W}_{xp}$ can be viewed as representing synapses in early visual cortex which connect neurons in V1→V2→...→IT pathway. $\mathbf{W}_{hp}$ can be viewed as synapses in the top-down pathway from frontal cortex early visual areas (Bar et al., 2006), and the bias term $\mathbf{b}_p$) represents tonic activations high-level sensory areas. $\mathbf{W}_{xi}$, $\mathbf{W}_{xf}$, and $\mathbf{W}_{xo}$ correspond to synapses into the basal ganglia from early visual areas, while $\mathbf{W}_{hi}$, $\mathbf{W}_{hf}$, and $\mathbf{W}_{ho}$ correspond to synapses into the basal ganglia from prefrontal areas supporting working memory (Purves et al., 2001, connections

[1]While the basal ganglia's connections to motor cortex are perhaps the best known (Purves et al., 2001, Figure 18.5), there is evidence that the basal ganglia also projects heavily to PFC in primates (Middleton and Strick, 2002).

shown in Figure 18.4). Bias terms $\mathbf{b}_o$, $\mathbf{b}_i$, and, $\mathbf{b}_f$ correspond to tonic activations in the basal ganglia.

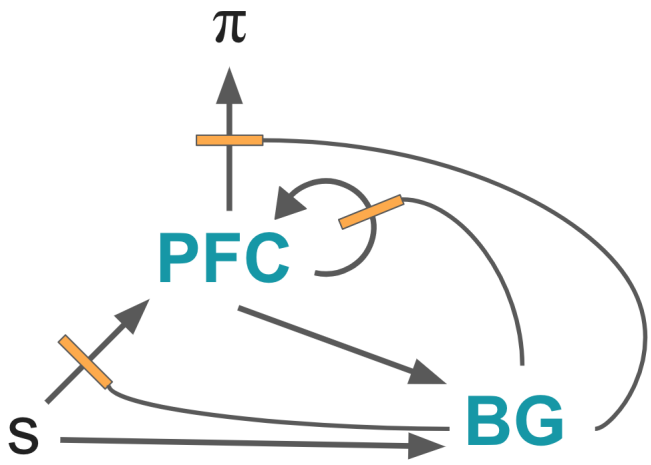


Figure 5.2: Summary diagram of input, output, and reallocation gates in the brain, useful for drawing an analogy from the LSTM's components to potential anatomical substrates. See main text for full description and connection to LSTM.

In the preceding paragraphs I developed a view of the LSTM working memory's components couched in the anatomy of the cortex and basal ganglia. In the upcoming section on reinstatement episodic memory in the brain, I will extend this view to hypothesize a biological implementation of reinstatement gating. To finish the discussion of the LSTM, I will first address the topic of backpropagation.

5.1.1 Backpropagation

The final and perhaps most controversial computational mechanism of the LSTM is the learning rule typically used to train it: backpropagation. Backpropagation works by recursively applying the chain rule to calculate the gradient of each unit and weight. The update for each weight for a given trajectory of experience is then computed as a learning rate multiplied by this gradient and added to the previous value of weight. This strategy applies in a straightforward manner to recurrent networks such as the LSTM by simply applying this recursive chain rule to each timestep

in a trajectory from the last to the first. A number of concerns have been raised regarding the feasibility of this algorithm in neural circuits, largely centering around the insufficiency of apparent mechanisms for propagating information in reverse, and the lack of evidence for other mechanisms by which gradient information might be computed and made available at the relevant synapse (for a more detailed summary of these issues see Bengio et al., 2015). Further, while training recurrent networks with backpropagation amounts to simply treating each timestep like separate layer of a feedforward network (resulting in an algorithm referred to as backpropagation-through-time), training recurrent networks with more biologically plausible learning rules poses additional difficulties (Marblestone et al., 2016).

This work used backpropagation, and more precisely, stochastic gradient descent (SGD) with a variety of tricks for determining the per-timestep, per-weight learning rate (e.g., Kingma and Ba, 2014) to optimize the weights of the EMRL architecture. In light of this fact, here are three different perspectives one could take when interpreting EMRL as a model of the brain

1. On one end of the spectrum, SGD could be viewed as a generic optimization procedure, not expected to bear any special resemblance to the optimization procedures used in the brain. In this view SGD exclusively serves the purpose of fitting the architecture's parameters to the environment, similarly to the way that an analytical optimization method would be used to derive optimal behavior given a set of bounds and a formal environment model (Oaksford and Chater, 1998). It is only intended to model learning rules at the synapse level insofar as it rationally maximizes reward.

2. A middle-ground view (and the one that I most often take) is agnostic about the degree to which SGD resembles the mechanisms which define synaptic learning in the brain, but holds that the *functional* effects, in terms of the manner of processing and representation of the trained system, of SGD and the brain's

learning rules are roughly the same. This remains a supposition; however, suggestive evidence comes from the fact that SGD does appear to induce representations similar to those found in the brain (Zipser and Andersen, 1988; Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014).

3. The other end of the spectrum holds that the brain uses learning rules that use approximate gradient estimates, typically invoking feedback connections that operate in a separate phase of a cycle from the feedforward connections (for review see section 2.2.2 of Marblestone et al., 2016). In this view, backpropagation can be taken somewhat more seriously a mechanistic model of the brain's learning rules.

The question of what synaptic learning rules the brain uses and how those relate to the SGD algorithms of deep learning appears to be under energetic scrutiny, so it may be that a consensus will soon emerge about the degree to which backpropagation mechanistically resembles the brain's learning rules. In the meantime, backpropagation may be reasonably viewed as a functional model of neural learning, or can more conservatively be taken exclusively as an expeditious method for optimization.

5.2 Hippocampal Episodic Memory and the Differential Neural Dictionary

5.2.1 Attractor Networks

EMRL's DND can be viewed as an experimentally convenient stand-in for an attractor network, of the kind hypothesized to implemented by circuitry in the hippocampus Rolls (2007) in support of episodic memory. An attractor network is a set of neurons, typically with excitatory recurrent interconnections, whose dynamics settle into a stable pattern Eliasmith (2007). Such networks have long been studied from the

perspectives of generalization, categorization, error correction, and – most relevant for our purposes – memory (Hopfield, 1982). Attractor memory networks have two key operations: storage and retrieval. Storage of an activity vector y into an attractor network is done using some kind of Hebbian learning mechanism such as

$$\delta w_{ij} = k \cdot y_i \cdot y_j. \tag{5.7}$$

where y_i and y_j are each units in the network, w_{ij} is the strength of the synapse between them, and k is the learning rate (Rolls, 2007). Retrieval can be accomplished in such a network by running the recurrent dynamics of the network using the trained weights

$$y_i^t = f(h_i, y_i^{t-1}) \tag{5.8}$$

$$h_i = \sum_j y_j w_{ij} \tag{5.9}$$

(Rolls, 2007). Intuitively, the storage procedure produces weights w which will reproduce activations y_i given activations at y_j when the retrieval formula is recursively applied. In this way, the network can retrieve stored patterns based on a partial cue. This type of memory system has been studied extensively as a model of both working memory in the PFC (Lansner et al., 2013; Wang, 2001, 2010) and episodic memory in the hippocampus (for review see Rolls, 2013a).

5.2.2 The Hebb-Marr Model of the Hippocampus

More specifically, the hippocampal region CA3 appears to operate as such an attractor network (Wills et al., 2005). Based on this region's recurrent, excitatory connectivity, Rolls (1991), McNaughton (1991) and others theorized that this region implements attractor learning and dynamics that support the hippocampus' role in

episodic storage and retrieval (as reviewed Rolls, 2013b). Modeling work has supported and developed this notion of CA3 as an autoassociative pattern completion using simulations of both spiking and spike-rate neurons and analytical analyses of storage capacity. Further modeling work has used CA3's attractor dynamics to explain place cell activity (reviewed by Hasselmo, 2011).

One difficulty with attractor networks is the potential for interference between memories, a problem that is exacerbated by correlation among the saved memories (Gluck and Myers, 1997). One way of addressing this issue is to ensure the patterns saved in the attractor network are as orthogonal as possible. These observations provide a functional explanation for a striking physiological property of the dentate gyrus, a hippocampal subregion that provides substantial afferent connections into CA3. Specifically, neural activity in the dentate gyrus is unusually sparse; that is, a very small proportion of the neurons in this region are active at any given time (O'Reilly and Munakata, 2000). This property makes the dentate gyrus well suited for *pattern separation*, that is, decorrelating representations on its input by mapping them to one or only a few units. Each unit of the dentate gyrus projects to cells in CA3 yielding distributed decorrelated representations of dentate gyrus' input.

Entorhinal cortex, a region which receives projections from sensory and association areas, projects to dentate gyrus, providing an interface between the sparse autoassociator in the hippocampus and the rest of the brain. In order for CA3 to induce reinstatement outside of the cortex, there must be a way for the input patterns to DG to be reinstated in entorhinal cortex. While the dentate gyrus–CA3 circuit's strategy of sparsifying inputs to yield decorrelated memories minimizes interference, it presents a problem for retrieval: the random representations of inputs in CA3 do not necessarily correlate with the pattern of entorhinal activity that produced it (O'reilly and McClelland, 1994). The connectivity of another hippocampal subregion, CA1, suggests it may provide a solution. CA1 has bidirectional connections to entorhinal

cortex and receives input from CA3. CA1 could thus solve the retrieval problem by learning a mapping from representations in CA3 to those in concurrently present in entorhinal cortex (O'reilly and McClelland, 1994). See Figure 5.3 for a depiction of relevant hippocampal anatomy and a summary diagram of the entorhinal-dentate-CA3-CA1-entorhinal loop.

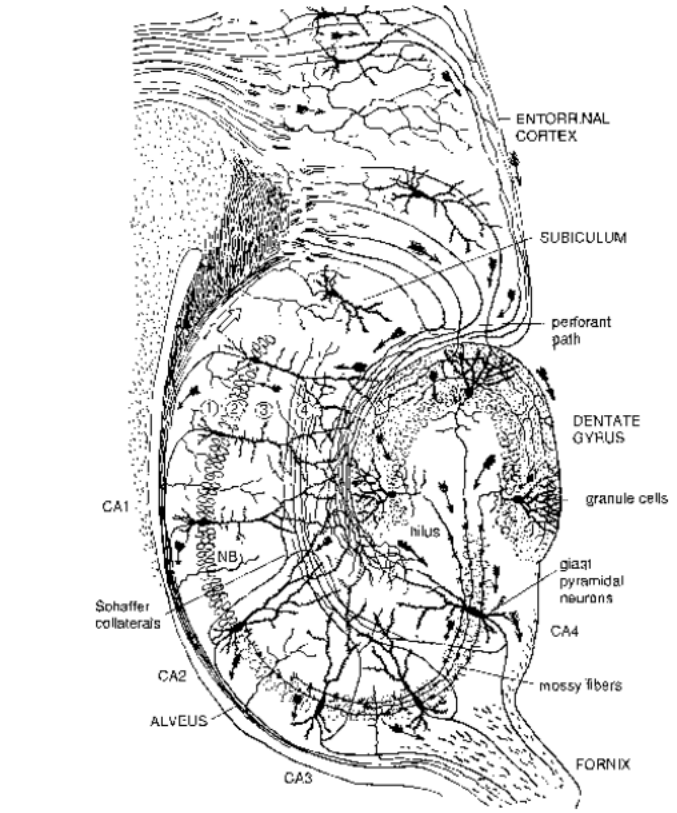


Figure 5.3: (Top) Reprinted from (O'Reilly and Munakata, 2000). (Bottom) Reprinted from (O'reilly and McClelland, 1994).

5.2.3 EMRL and Hebb-Marr

This theory of hippocampal function, wherein the dentate gyrus, CA3, and CA1 together provide interference-minimized attractor based storage and retrieval of patterns in entorhinal cortex, is often referred to as the Hebb-Marr model, and is dominant in neuroscientific thinking about the function and physiology of the hippocampus (e.g., Rolls, 2007, 2013b, 1991, 2013a; O'reilly and McClelland, 1994; McNaughton, 1991; Treves and Rolls, 1992). How does the DND relate to this model, and is it possible to reasonable to anticipate that behaviors would be reproduced by an implementation of the Hebb-Marr model?

The DND provides the storage and pattern retrieval from a partial cue that the CA3 attractor network of the Hebb-Marr model provides. Rather than storing patterns in a set of Hebbian-learned weights, the DND simply appends patterns to separate slots in an array. To retrieve, the DND computes the similarity between the retrieval cue and all of the patterns stored in the array, using a similarity metric such as a dot product or cosine distance. The DND returns the pattern that is most similar under this distance metric. This is functionally equivalent to the attractor network from Equation 5.7, which retrieves the pattern stored in memory that has the largest dot product with the query cue (Rolls, 2010). The DND avoids catastrophic interference by simply storing each entry in separate slots, obviating the need for sparse projections and pattern separation provided by the dentate gyrus in the Hebb-Marr model. Likewise, this removes the need for the invertible mapping from the stored pattern to the cortical activities that produced them; because the cortical activities are stored in their usual format, the retrieved patterns can be be reinstated directly.

This mapping between the DND and the Hebb-Marr model suggests that behavior that results generally from the storage and interference-free, similarity-maximizing retrieval of patterns can be reasonably expected to be invariant between the two systems. In contrast, behavioral variation due to idiosyncrasies of sparsified intermediate

representations, Hebbian learning dynamics, or errors in converting pattern-separated representation to their original form will be captured by a Hebb-Marr model, but not by EMRL. A corollary of this reasoning is that anything EMRL can do, a Hebb-Marr should be able to do as well, assuming effective sparsification and reconstitution of saved patterns. Thus the behavior studied in the simulations in this thesis — e.g., the retrieval of partial bandit solutions, and episodic model-based learning on the two-step task — are like to be representative of the behavior we could expect from an implementation of a Hebb-Marr model. Future simulation work will be needed to confirm this reasoning.

A question that naturally arises at this point is why should we use EMRL in simulation rather than a Hebb-Marr model, if EMRL is expected to capture fewer details of human behavior? The answer is threefold. First, the DND is easier to implement and successfully train, precisely because concerns about sparsification, catastrophic interference, and effective invertible mappings are obviated. Second, the DND can scale to much larger problems that we could approach with a Hebb-Marr model, in part thanks to synergy with interest in machine learning in such algorithms (Pritzel et al., 2017; Kaiser et al., 2017). Without the ability to scale these simulations to complex environments such as 3D first-person tasks (Beattie et al., 2016) and Starcraft (Vinyals et al., 2017), confidence that our findings could scale to the level of human intelligence would be greatly diminished. Finally, and perhaps most importantly, simplified systems like the DND render transparent the *sufficient conditions* for behaviors of interest. Had the simulations in this thesis been carried out using a more biologically detailed model, it may have been less obvious that the observed effects were not reliant on the details.

In sum, the DND can be viewed as a convenient, minimal stand-in for more biologically detailed models of hippocampal function. An appreciation of the functional

similarities between these types of models will hopefully enable us to capitalize maximally on the strengths of both.

5.3 Gated Reinstatement: Connecting Working and Episodic Memory

This chapter so far has discussed the mechanisms by which and anatomical locations in which the algorithms carried out by each major component of EMRL, i.e. the LSTM and the DND, might be implemented in the brain. This section discusses how episodic and working memory in the brain might be connected in a manner analogous to EMRL's gated reinstatement.

EMRL suggests that the similar mechanisms to those that support input gating to working memory (Frank et al., 2001) might support gating of hippocampally reinstated working memory activations. Fortuitously, empirical work has investigated the possibility of gating mechanisms operating on the hippocampal-prefrontal pathway. Floresco and Grace (2003) stimulated neurons in fornix-fimbria pathway, a major output track from the hippocampus, evoking firing in PFC neurons. Intriguingly, they showed that concurrently stimulating neurons in the medial dorsal thalamus inhibited the firing of the PFC neurons in response to fornix-fimbria stimulation. In another population of thalamic neurons, stimulation induced short term facilitation of fornix-fimbria firing of some prefrontal neurons, and inhibition of others. Floresco and Grace further showed that concurrent stimulation of the ventral tegmental area modulated fornix-fimbria induced prefrontal firing. The valence of this modulation depended on the frequency and relative timing of stimulation. While these results are somewhat too complex to relate cleanly to the input gating mechanism of Frank et al. (2001), it is striking that thalamic activation should modulate the hippocampal-PFC pathway as it does for the sensory-PFC pathway in that model. In Frank et

al.'s model, gating applies to rostrocaudal propagation of activity from sensory and association cortex into PFC. While this is sensible for input gating, in does not apply to the direct reinstatement of prefrontal activities. Together these facts suggest that while the input and reinstatement gating mechanisms may be similar, and may both rely on thalamic modulation, there exist important differences between them.

Anderson et al. (2016) put forth two other hypothesis about the mechanisms underlying what they call "retrieval suppression". The first is that inhibitory projections from anterior cingulate to entorhinal cortex block activity from entering the hippocampal retrieval loop, thereby preventing retrieval. In contrast to the mechanisms at work in EMRL and those apparently at work in Floresco and Grace's experiments, this mechanism blocks recall at the point of memory querying the memory, rather that at the point of reinstatement. This hypothesis is based on observations about connectivity, and future work will be needed to see whether the known connectivity functions in this way. Anderson and colleague's second hypothesis is that connections from anterior cingulate cortex to thalamus to CA1 might gate CA3 representations retrieved from episodic memory from producing reinstatement in entorhinal cortex and beyond. This proposal as well is based on the observation of inhibitory connectivity, so determining whether these connections function as reinstatement gates will require future work.

In summary, there is evidence for reinstatement gating and interest on the part of the experimental neuroscience community in a developing a fuller mechanistic understanding of this process. As for working memory gating, the exact mechanisms are at this contested, and it is possible that multiple mechanisms are at work.

5.3.1 Modeling Directions: Implementational-Level Analogues to EMRL

This chapter's discussions so far set up two open questions that more biologically detailed instantiations of EMRL could help to solve. First, what mechanisms underlie the retrieval suppression reviewed by Anderson et al. (2016)? To answer this question, a model that is highly specific at the systems level and moderately detailed at the cellular level (enough to e.g. distinguish excitatory and inhibitory populations) in the spirit of Frank et al. (2001) and O'reilly and McClelland (1994) could be very useful for making experimental predictions. As a matter of fact, the conjunction of Frank et al.'s model of prefrontal working memory gated by the double-inhibitory basal-ganglia–thalamic circuit could be combined with O'Reilly and McClelland's interference-minimized hippocampal attractor network via one or more of the reinstatement mechanisms discussed in the previous section. The intersection of these two models contains the prefrontal, entorhinal, CA1, and thalamic areas necessary to instantiate both of Anderson and colleagues' (2016) hypotheses[2] by adding additional connections. Such a family of models may be able to generate differential predictions for each gating hypothesis, spurring the process of experimentally refining our understanding of reinstatement gating in humans.

The second open question has two parts — what additional challenges arise when EMRL is implemented using an attractor network as the episodic memory and a continuous-time recurrent network as the working memory, and what additional behavioral variation can be explained by the solutions to these challenges? For example, it is clear that such an attractor memory system would require separate phases for storage and retrieval, and these phases may result in the behavioral effects measured by Duncan and Shohamy (2016, see Chapter 6). Issues like this one, that turn on

[2]Incorporating elements of models of dopaminergic gating of working memory, such as Braver and Cohen (2000), could be helpful for integrating the ventral tegmental area's apparent role in hippocampus-prefrontal gating (Floresco and Grace, 2003) with the role of thalamic circuits.

algorithmic properties such as those inherent to attractor networks in general may be more clearly demonstrated using a model less elaborate than an O'Reilly and McClelland-style Hebb-Marr model.

Consider the following alternative, made up of spike-rate (connectionist) units: a network of nodes trained using attractor learning weight update from Equation 5.7 is connected via a subset of its units to a network of units with the same type of recurrent connectivity as the attractor network (shown in Equation 5.8) but which output a policy and whose weights are trained to maximize a reward objective. Gates could be implemented by an additional set of units that receive input from the latter set of units and connect multiplicatively to outputs of the attractor network. In order to solve the tasks from Chapters 3 and 4, this model would require some method for transitioning from storage mode to retrieval mode, and perhaps solutions to a variety of other problems.

In addition to helping to identify difficulties lurking in the rough equivalence I have drawn between the LSTM with DND-based reinstatement and reinstatement via attractor networks, such a model would be well positioned to address known findings. For example, Wimmer et al. (2014) showed an inverse relationship between incremental and episodic learning and that functional connectivity between striatum and hippocampus co-occurred with poor incremental learning. A model like the one above could engage with such a finding by modulating the Hebbian learning in the attractor network by the RPE signal from the reward objective function (see Chapter 5 for more detailed discussion). Indeed, this type of modulation may be an important aspect of a model at this level of detail; past empirical and theoretical work suggests that such modulation may be an important aspect of hippocampal learning (Wimmer and Shohamy, 2011; Jang et al., 2018). Similarly, Duncan and Shohamy's (2016) finding would be fully within the scope of such a model. Such a system would provide a minimal model for addressing these issues, in contrast

to the additional systems-level complexity of the Hebb-Marr model and the cellular granularity of models like Frank et al. (2001). This simplicity would enable clearer interpretation and might provide insights more accessible to a broad audience.

In summary, this chapter discussed how and where the basic components of EMRL might be implemented in the brain, and outlined two more biologically detailed analogues to EMRL. This opens the door for future work to simulate and test the predictions of these models, to further our understanding of the interactions between working and episodic memory in humans.

Chapter 6

Empirical Connections

A number of recent studies have investigated, behaviorally and neurally, the interactions between episodic memory and the traditionally studied mechanisms behind reward-based learning and decision making, revealing a number of surprising findings. To be taken seriously as a candidate model of reward-based learning, episodic meta-RL (EMRL) should be able to explain these findings. In this chapter I discuss a set of especially important and relevant experiments, describing their key results, and how EMRL would account for them, or could be modified to do so. To account for several of these findings, EMRL would need to be modified, or a more biologically detailed instantiation of EMRL will need to be invoked. In making these modification recommendations, I have endeavored to avoid highly specific, bespoke changes for the results at hand; instead the modifications should have general utility for increasing the competence of artificial agents and for modeling human cognition. Encouragingly, a number of these findings call for the same modifications. The next several sections will deal with each relevant study in turn, and will be followed by sections synthesizing the implications of these experimental findings for EMRL. The chapter concludes with prospective thoughts about the potential for modelling a large set

of these findings simultaneously following the connectionist rational analysis (CRA) approach.

6.1 "Reminders of past choices bias decisions for reward in humans" – Bornstein et al.

6.1.1 Experiment 1

Bornstein et al. (2017) provide a clear minimal demonstration of episodic learning during decision making in humans. First, they reanalyzed data from a seminal experiment by Daw et al. (2006), wherein participants performed a multi-armed wandering bandit task. They found that an episodic sampling model fits the behavior, RPEs (measured by fMRI), and chosen value (measured by fMRI) better than the incremental average model used in the original study. Even this very simple experiment poses challenges to EMRL as it was used in Chapters 4 and 5, challenges that will lead to two of the major themes of this chapter.

The first challenge is that in this experiment, there is no explicit stimulus or cue, just a set of arms to pull on each timestep. As such, there is nothing with which to query the episodic memory, or in other words, nothing to put in the keys column of the differentiable neural dictionary (DND). This is akin to an important property of human experience — sensory input does not come pre-divided into episodic query keys and other working memory input. As such, it is necessary to parse the stream of sensory input in order to glean the appropriate information with which to query episodic memory.

The second challenge is that in this task, there's no *reward incentive* to use an episodic sampling strategy. As such, there is no guarantee the agent would use the learn from this task to use episodic sampling rather than incremental averaging —

the EMRL has the architectural capacity to do either, at least as those strategies are defined in this task by Bornstein et al. (2017). This stands in contrast to the episodic two-step task of Chapter 5, in which there was a small but significant reward incentive to performing episodic RL.

Of course, there's no reward incentive for humans to use episodic sampling in Daw et al.'s (2006) task either, so why do humans do it? One plausible answer is that humans use episodic sampling because there *is* a reward incentive for episodic sampling in the environments that humans inhabit outside of the lab, and they bring inductive biases built from that experience into the task. Indeed, natural environments often pose recurring underlying states that determine the efficacy various actions. Plonsky et al. (2015) formalize this observation into environment models consisting of latent Markov chains, each state of which determines the reward probabilities of a bandit problem. Plonsky and colleagues showed that sampling models that query based on the recent history of actions and rewards reproduce otherwise challenging aspects of human behavior in decision making tasks.

Putting these two observations together, we see that to explain this finding, EMRL needs (1) an architectural improvement that enables it to query episodic memory effectively when all it receives from the environment is sequences of actions and rewards and (2) a training distribution that captures important intuitions about the natural world and in effect provides reward incentive for learning to execute episodic RL.

What architectural change would enable EMRL to retrieve information by match between the recent action-reward history and previously experienced action-reward histories? Since the LSTM hidden state should encode these quantities, an appealing option is to query using the hidden state itself as the key. We'll see throughout this chapter (as well as at the end of Chapter 4) that moving to such an autoassociative retrieval mechanism is an essential next step for research with EMRL.

Combining such a working memory query mechanism with a task distribution like that of Plonsky et al. (2015) would put EMRL in a good position to address this first experiment from Bornstein et al. (2017)[1].

6.1.2 Experiment 2

In their second experiment, Bornstein and colleagues ran a modified version of Experiment 1 wherein unique images were shown during some trials and later presented again in a paid recall test interspersed with more choices. T-tests on arm selection probability indicated a significant effect of reward during initial presentation on choice. Further, a sampling model that took the cued trial into account fit the data better than the incremental model.

EMRL as it was used in Chapters 4 and 5 is unlikely to capture this effect. The obvious way to model this experiment would be to embed the stimuli and use them as query keys. Such a system would learn to correctly classify stimuli as previously seen or unseen (solving the recall test) because the query mechanism would find a match in the DND if the stimulus was presented before, otherwise nothing would be retrieved. In the case that a match was found, the associated hidden state would be reinstated. This reinstated hidden state could bias the agent's decision making as observed in human behavior; however, this bias would lead to decreased reward, because the recent experience is better correlated with the current reward parameters than the experience associated with the original presentation of the stimulus. As a result, the agent would probably learn to ignore the information in the retrieved hidden state, either via gating or by some other mechanism.

[1]It is worth pointing out that it remains an open empirical question whether MRL – that is, the meta-learning LSTM *without* episodic memory – would be better fit by the episodic sampling or incremental model from Bornstein et al. (2017). It is plausible that the episodic sampling model would fit quite well because it only samples recent past episodes within a window that could easily be stored in the LSTM hidden state. In the above analysis, I worked from the assumption that the LSTM-only architecture trained directly on this task would not capture the sampling effect, but future simulation work will be needed to verify this assumption.

A more certain way to capture this finding would be to train EMRL on a task distribution that incorporates our intuitions about why human behavior should be biased toward actions that were effective in situations with similar appearance. In the real world, visual observations provide cues to underlying structure. For instance, restaurants with certain types of decor tend to excel in certain types of dishes. Incorporating this property into a training distribution might well be enough to capture this effect. Concretely, future work could take the bandits with underlying MDPs of (Plonsky et al., 2015) and add observable features that indicate (perhaps with some noise) the underlying state. Training on an appropriate distribution of this type should put EMRL in a good position to reproduce the effect in the second experiment of Bornstein et al. (2017).

To cover both experiments from Bornstein et al. (2017) with a single architecture, it will be necessary to reproduce Experiment 2 with an architecture that doesn't require a pre-parsed cue, as is necessary in Experiment 1. This can be accomplished for Experiment 2 by feeding the stimulus images into the LSTM, then querying using a function of the LSTM hidden state. At the outset of training, there would be some degree of match between the LSTM state during the original presentation of the stimulus and during the recall test simply by nature of matching random projections of stimulus. The reward signal from the recall task propagating through the DND's differentiable pathway may then be enough to train query embeddings that reliably retrieve the correct hidden state. This as far as I know would be a novel demonstration, but there are no obvious barriers to success.

6.2 "Reinstated episodic context guides sampling-based decisions for reward" – Bornstein and Norman

Bornstein and Norman (2017) conducted an experiment in which participants performed a wandering bandits task with "context" images that lasted for many consecutive trials and "cue" images that lasted for one trial. Shortly after the beginning of the context, the arm probabilities shifted dramatically, enabling a dissociation of the effects of cued and context episodes. In a final block of trials without contexts or cues, recognition memory questions were interspersed among the bandit trials. Bornstein and Norman found that the episodes associated with the context influenced behavior in the final block significantly when the cue images from the beginning of the context were re-shown. The astounding finding is that the context episodes after the cued episodes influenced behavior more than the cued episodes themselves.

EMRL as used in Chapters 4 and 5 will not reproduce this finding. The DND would, during recognition probes, return hidden states present when the stimulus was first encountered. Under the right training circumstances, this could bias behavior in subsequent trials towards those that were taken in the presence of the cue, as observed in humans. However, due to the clever design of this experiment, the retrieved hidden states would contain no information about the context trials other the cue trial; therefore, this architecture would not reproduce the context influence effect.

As for the previous study, for EMRL to account for this finding, it will be necessary to query using the hidden state instead of a separate perceptual cue. Assuming that the trial stimuli and recognition probes are provided as input to the LSTM, and that the reward from the recognition probes induces an effective query mechanism (as in Experiment 2 of the last section), then the LSTM hidden state present during the original presentation of the stimulus will be reinstated during the recognition probe.

The retrieved hidden state will encode the context, as well as the cue. As a result, on the next timestep the query key will contain an encoding of the context, making it likely that other states from the same context are retrieved[2].

Note that here, like for Experiment 2 of Bornstein et al. (2017), there is no correlation between the bandit parameters during the cued context and those after the probe trial, so in this setting EMRL would learn to ignore information from the episodic memory when making decisions. Similarly to that experiment, it would be necessary to train EMRL on an appropriate task distribution in which there was a reward correlation (in effect modeling a property of the natural world) in order to reproduce this effect.

Additionally, Bornstein and Norman (2017) found that increased pattern reinstatement (as measured by a scene classifier) co-occurred with increased effect of cued context on choice behavior. An analogous effect to look for in EMRL would be that the degree of cell reinstatement (either measured by the gate values or by decodability) should correlate positively with the strength of the effect of the context trials on the agent's behavior.

6.3 "Episodic Memory Encoding Interferes with Reward Learning and Decreases Striatal Prediction Errors" – Wimmer et al.

In a study by Wimmer et al. (2014), participants performed a 2-armed bandits task, where on each trial, each arm was paired with a unique object image. Each arm was also paired consistently with a color, and the reward probabilities associated with the

[2]To get this context effect without waiting for the next timestep, the architecture could query the DND multiple times for per environment timestep. Effectively, this would mean running the LSTM with reinstatement for multiple timesteps per each environment timestep. Such a scheme would come closer to continuous time in which biological neural circuits operate.

colors drifted over time. The next day, participants performed a surprise recognition memory test for the object images. The authors find evidence that the effect of a trial on subsequent choice (as indicated by a model fit) is weaker when episodic memory encoding is stronger (as measured by the recognition test).

This effect can not be (in any straightforward manner) accounted for by EMRL as it is used in Chapters 3 and 4, where the stimuli embeddings are used the query keys. This is because the recognition test would simply involve embedding the stimuli and asking whether there is a match in the DND, and the outcome of this query would be totally independent of the associated LSTM state and its encoding (or lack thereof) the trial outcome.

In contrast, if EMRL queries using the hidden state itself, a potential explanation emerges. If the LSTM has limited capacity then it can store either the outcome of trial *or* the stimulus. In such a regime, the trial outcome might have a smaller effect on subsequent behavior when episodic encoding is better because the outcome is not encoded as clearly in the hidden state. This explanation is problematic, because I wouldn't expect it to hold if the hidden state was sufficiently large. Further, I expect that gradient learning during the recall phase would push recall performance to ceiling in versions of this task with trial blocks of reasonable length.

Additionally, this explanation does not accord with the fMRI findings Wimmer and colleagues describe: that the striatal RPE signal was weaker when episodic memory was stronger, and that weakened RPE signals co-occurred with increased functional connectivity between striatum and hippocampus. EMRL's RPE is not influenced by, and does not immediately influence the episodic memory storage process. As discussed in Chapter 5, we take the DND as a convenient stand-in for an attractor network whose dynamics are shaped by Hebbian learning, the learning rate of which could be modulated by dopamine (Wimmer and Shohamy, 2011). When viewed at this lower level of abstraction, EMRL is better able to engage with this finding. It

is tempting to suggest that dopamine (or as Wimmer et al. (2014) put it, "striatal processing") is a limited resource that can either be used to modulate Hebbian learning in the hippocampus *or* to effect learning in the prefrontal network. However, in MRL, incremental learning is supposed to be carried in activations, without the need for RPEs, so this explanation too falls somewhat flat.

In short, while EMRL engages with these findings, it does not provide a totally satisfactory account. It could be that more biological detail, separating the roles of striatum and PFC for instance, is necessary to convincingly explain these results. Additionally, it may be worth waiting to see whether this dopamine-mediated trade-off between striatal and hippocampal learning is replicated before constraining the theory too heavily.

6.4 "Memory States Influence Value-Based Decisions" – Duncan and Shohamy

In their first experiment, Duncan and Shohamy (2016) tasked participants to select between cards from two decks. On each trial, participants saw the face of one card from each deck. On the back of each card was a hidden reward amount, which the participant would receive after choosing that card. Some of the cards were presented multiple times during the experiment, and participants could make more money by remembering the payout amount of previously seen cards. Choices were carried out in the presence of "context" images, which also reoccurred throughout the experiment. On the first presentation of a context it is considered a "novel", and on later presentations it is considered a "familiar" context.

Duncan and Shohamy found that the influence of the previous presentation of a card was greater in familiar contexts than in novel contexts. A tempting interpretation is that the familiar context images sometimes match the context present during the

previous card presentation, leading to greater overall perceptual similarity, leading to increased retrieval strength. However, Duncan and Shohamy ruled this out by ensuring that arm images never re-occurred with the same context image. Thus, this effect is about familiarity *per se*, not the retrieval of specific memory traces based the familiar context. In Experiment 3 the authors ran the same experiment with a larger sample size and showed that arm images seen in novel contexts had greater influence on later decisions about those arm images than arm images seen in familiar contexts[3]. Duncan and Shohamy explain these effects as results of alternating hippocampal states, one for encoding, and one for retrieval. Under their explanation, the familiar contexts induce a lingering[4]retrieval state, while the novel contexts induce an encoding state.

There are two ways to address these findings in the EMRL framework. The first is to argue that, in contrast to Duncan and Shohamy's theory, hippocampal states are not involved. Instead, retrieval effects are an adaptive response to an environment wherein familiar contexts, it generally pays off to pay attention to memories from similar states, while in novel contexts, it generally pays off to ignore long past memories and rely on recent outcomes. Similarly, the storage effect may be sufficiently explained as an adaptive response to an environment wherein novel contexts tend to co-occur with important novel information that is not present in semantic memory, indicating increased utility in episodic storage. Pursuing this explanation with EMRL would involve constructing a training distribution in which these properties hold, then testing whether the generic autoassociative architecture described for the previous experiments in this chapter reproduces these effects.

The alternative route to explaining these findings with EMRL is to say that indeed there are two separate hippocampal states, and that the DND implementation

[4]This effect was present but not statistically significant in the first experiment with the smaller sample size.

[4]Experiment 2 shows that the retrieval effect obtains when the context images are displayed before and not during the choice.

abstracts over this. In order to engage with this finding, it is necessary to invoke the more biologically rich analogue to the DND (described in Chapter 5), wherein storage is carried out by Hebbian learning and retrieval is carried out by attractor dynamics. In this modeling setting it *is* necessary for there to be two retrieval states, and we would be in a position develop the theory further with conditions (such as novelty and familiarity) that influence the hippocampal state.

These two routes to addressing these findings are not mutually exclusive; indeed, it may be that the most effective theory will combine the computational level perspective and the implementational detail. In an EMRL-style model this could be done by implementing the two hippocampal states, where the function that swaps between states is be learned by interaction with a training distribution in which novelty predicts utility for storage and familiarity predicts utility for retrieval.

6.5 "Refresh my memory: Episodic memory reinstatements intrude on working memory maintenance" – Hoskin et al.

In this chapter's first foray beyond bandit-like decision making tasks, I will now discuss a list memory study by Hoskin et al. (2018), the results of which suggest that episodic memory reinstates working memory representations in. In their first experiment Hoskin and colleagues had participants carry out a no-delay recall task of sets of words paired with context images. After learning word associations through this task, participants carried out another set of recall trials, this time with an intervening delay of 18 seconds that included a variable length distracter task. Hoskin and colleagues found that participants made more intrusion errors the longer the distracter task. Further, participants were more likely than expected by chance to incorrectly

answer with words that were paired with the to-be-recalled word list during the first stage task. Hoskin and colleagues take this intrusion effect as evidence that episodic memory comes to working memory's aid in the face of distracting task demands, at the cost of reinstating related words not in the to-be-remembered word list.

The EMRL architecture is in a position to simulate this task, if its policy and value layers are replaced by a softmax layer with an element for each output vocabulary item and the policy gradient RL loss function is replaced by a supervised loss function such as negative log likelihood. This setup is commonly used used in language modeling (Mikolov et al., 2010), machine translation (Sutskever et al., 2014), and has been used to model human list memory (Botvinick and Plaut, 2006). In this framework, both stages of this experiment can be treated as sequence autoencoding tasks, wherein the words in the list are presented as inputs to the LSTM, one per timestep. Once the full list has been presented, the LSTM's targets are the list items in order, one per timestep. In this manner, the LSTM learns to encode the list items in its hidden state (Botvinick and Plaut, 2006; Kiros et al., 2015).

If the architecture fills the DND with each LSTM state calculated during the association-learning phase, then during the second phase, if the working memory capacity is overtaxed, the system could learn to store just enough of the to-be-recalled list to serve as a query key into episodic memory to retrieve associations. This strategy will be much more effective than chance guessing, but will have the side effect of increasing the number of intrusions from the associated set of words. Future simulation work would be necessary to work out exactly the conditions under which this might happen. An important condition would be the ratio between the working memory capacity and length of the list to remembered, because if the working memory is large enough, the system can simply ignore the episodic memory.

Without further elaboration, this story fails to explain the fMRI results from the third experiment in Hoskin et al. (2018). In a paradigm similar to Experiment 1

described above, Hoskin and colleagues found evidence for pattern reinstatement of the associated word lists *during the delay phase*. Further, the amount of evidence predicted performance on some trials. The setup above would account for reinstatement during the scene associated with the first-stage word lists during recall but not necessarily during the delay phase. The only way that EMRL might reproduce this effect is if the shortest description query key is an encoding of the context image itself, so that the LSTM is incentivized to query the DND during the presentation of the to-be-remembered list, infers the associated image, and stores that encoding through the delay in order to query with it at the end. Without further simulation work this is essentially a just-so story, and I am not at all confident that gradient learning would lead to such an effect.

In my view, a much more appealing direction for explaining this finding is one suggestion by Hoskin et al. (2018) that this reinstatement may be a mechanism for forward planning: "the ability to interact with working memory may be an adaptive feature of resting-state replay from episodic memory in other words, it may not just sustain, but also transform working memory representations, by integrating information in working memory with that from recent events". This suggests that a task distribution wherein interactions between ongoing working memory and episodic memory are adaptive might be the most appropriate way to explain this finding in the EMRL framework.

In sum, while the basic mechanics of EMRL are well positioned to simulate the experiments from Hoskin et al. (2018) and a training distribution that closely resembles the task itself could lead to the demonstrated behavioral effects under the right conditions, it is unlikely that EMRL in its current form would convincingly reproduce the fMRI finding. A promising direction to improve EMRL to explain this finding would be to hypothesize an adaptive role of reinstatement — such as for one-shot

forward planning — in relevant settings outside of this task, then to design a training distribution that captures important features of those settings.

6.6 "Multiple Processes in Prospective Memory Retrieval: Factors Determining Monitoring Versus Spontaneous Retrieval" – Einstein et al.

One relatively old thread of work that bears on interactions between episodic and working memory investigates the mechanisms underlying prospective memory (Hoskin et al., 2018; Cohen and O'Reilly, 1996). In light of this relevance, this chapter discusses how EMRL might apply to the tasks studied in this work, focusing on a representative example study by Einstein et al. (2005). In a series of experiments, participants were shown noun/category pairs and asked to classify whether the noun was a member of a category. Einstein and colleagues refer to this as the ongoing (OG) task. Simultaneously, participants were asked to respond whenever they saw a certain string of characters. This task was referred to as the prospective memory (PM) task. In a set of 5 experiments, Einstein and colleagues demonstrate that under certain conditions, people are able to perform the PM task without experiencing degradation in their performance on the OG task. This is taken as evidence that in addition to solving the PM task by monitoring for the PM cue, people are able to "spontaneously" retrieve the PM response criteria when the cue is presented. In other words, people can do the PM task without devoting a significant portion of working memory capacity to monitoring for the cue[5].

[5]Einstein et al. (2005) experimentally delineate the conditions under which people use monitoring, spontaneous retrieval, and combinations of both. In this discussion, I focus on how EMRL might

Such spontaneous retrieval could be a useful (and to my knowledge, undeveloped) capability for deep RL agents. Further, spontaneous retrieval is likely a result of episodic/working memory interaction (Cohen and O'Reilly, 1996), and as such should be explained by an aspiring theory of episodic and working memory. How could EMRL accomplish such spontaneous retrieval? It is conceivable that, when trained directly on a PM+OG task, for example a challenging n-back task paired with a PM target, EMRL would learn a spontaneous retrieval solution. Such a solution could involve, for instance, encoding the PM cue channel during the OG task with just enough precision for the query function to retrieve the target recorded at the beginning of the episode.

While this strategy would use less working memory capacity than a strategy that fully encodes the cue, some working memory space would be occupied. This would mean that if the OG task were challenging enough, adding the PM would produce some deficit. An alternative modeling direction would be to add a direct pathway from the perceptual representation c_{in} to the query function, so that the PM cue could bypass working memory altogether. This would alleviate the need for EMRL to encode the PM cue in working memory at, enabling solutions to the PM task that do not affect the OG task performance, no matter how challenging the OG task. Simulation work will be needed to confirm these predictions.

It's worth noting that both of these modeling possibilities accords with the reflexive-associative theory Einstein et al. summarize, wherein "an association [is formed] between the target cue and the intended action ... when the target event is encountered, an automatic associative system (like the hippocampal system ...) delivers the intended action to consciousness". EMRL would essentially provide a concrete implementation of this idea.

---

accomplish the important feat of spontaneous retrieval at all, setting aside considerations of the conditions that induce monitoring vs spontaneous retrieval.

Of further interest, such OG+PM tasks could be useful for developing deep RL agents that are able to make event conditional plans, then free working memory to carry out intervening tasks, executing the planned actions when the time is right. Such capabilities would be especially useful in complex domains that require flexible multi-tasking, such as Starcraft (Vinyals et al., 2017).

6.7 General Discussion

A number of themes emerge from the chapter's reflections on empirical findings and EMRL. Perhaps the most salient is the need to move to querying using hidden states rather than separate perceptual keys. In fact, in the case of all six of the studies discussed, this modification is necessary. Further, as discussed at the end of Chapter 3, querying with (functions of) the hidden state addresses an outstanding problem for deep RL agents as well — environments where the information needed to query effectively is distributed across multiple timesteps. This change would also come closer to the kind of autoassociative memory traditionally attributed to the hippocampus (Marr, 1971). As discussed in the section on spontaneous retrieval, it may be important in some settings to query episodic memory with perceptual representations before they reach working memory; a complete model may require querying from both working memory and perceptual cortices.

Another theme that stands out across multiple of these studies is the importance of the environment model. In four out the six studies discussed, I described task distributions other the experimental task itself as key ingredients in the process of reproducing the human behavior. These task distributions were motivated by the intuition that the observed behavior resulted in part from learning in the real world, and that the relevant features of the real work could be captured in a formal environment model.

The broad range of learning behaviors that a given architecture might produce as a result of various training distributions raises an important question about the utility of connectionist rational analysis (CRA; i.e., the task-distribution oriented flavor of modeling this thesis pursues) as a theoretical paradigm. Specifically, is a given architectural model falsifiable? A reasonable concern is that there are too many "degrees of freedom" so to speak, in the environment model, such that any new finding can be accommodated by manipulating the task distribution. This concern echoes the criticism often levelled at the Bayesian modeling paradigm, that manipulations of the prior can enable a theory to cover too broad a range of behavior (Bowers and Davis, 2012).

While this observation warrants caution, there seem to be a number of safeguards in CRA that prevent architectures from being unfalsifiable. First, certain architectures will never capture certain phenomena, at least not without wildly implausible training distributions. For example, the episodic sampling discussed by Bornstein et al. (2017) could never be accounted for by the EMRL agent used in Chapters 3 and 4, because the experiment doesn't provide perceptual cues with which to query the memory. Second, environment models should be designed to capture properties believed to inhere in the natural (which might ideally be chosen *a priori*). This measure of face validity provides a useful constraint on the environment models. Finally, if the emergence of behavior depends on environmental features that are too specific it is less likely that these behaviors will emerge during simulation.

The last theme that stretches across the discussed studies is the importance of an implementation-level analogue to the LSTM and DND-based EMRL used in previous chapters. In grappling with the findings of Wimmer et al. (2014) and explaining those of Duncan and Shohamy (2016), it was important to invoke the machinations of attractor dynamics and (dopaminergically modulated) hippocampal Hebbian learning. Without a mapping from the DND's storage and retrieval operations to these biolog-

ically feasible processes, EMRL would lose considerable explanatory power. Indeed, this may be one of the major strengths of the modeling approach followed by MRL and EMRL: it can incorporate both computational level and implementational level constraints.

One other point of interest with regard to empirical connections is the possibility of reproducing multiple experiments from diverse sources with a single architecture and training distribution; i.e., train a single agent which reproduces multiple findings without further gradient training. In this chapter a number of experiments were discussed that have essentially the same interface: the experiments of Bornstein et al. (2017), Bornstein and Norman (2017), Wimmer et al. (2014), and Duncan and Shohamy (2016) are all essentially multi-choice tasks with (in some cases) cue stimuli and context stimuli. Future work may train agents on task distributions consisting of multi-choice (bandit-like) tasks with random underlying MDPs, as in (Plonsky et al., 2015, Figure 3), where sometimes cue and/or context stimuli indicate the otherwise unobserved state of the MDP. The intention would be to then test the agent on the tasks from the aforementioned studies, and assess whether the trained agent reproduces some or all of the findings. Future work will be necessary to explore this highly speculative possibility; if successful it could provide a very simple explanation in terms (ideally a small number) of environment and architectural features for a broad range of findings.

Chapter 7

Broader Connections

Because of its stance at the intersection of working memory, episodic memory, and decision making, meta-RL with reinstatement relates to a broad range of other neuroscientific theories and literatures. This chapter discusses implications of a variety of theories and findings for EMRL and vice versa, suggesting along the way future simulation work that leverages EMRL to better understand behavior or applies neuroscientific ideas to increase EMRL's capabilities.

7.0.1 Temporal Context Model

Originally introduced to explain the recency and contiguity effects in free recall, the temporal context model (TCM) posits that people maintain a representation of "context", which drifts over time in response to new inputs (Howard and Kahana, 2002, 1999). When an item is recalled, the state of the context representation is reset to the state present during encoding of the item. The retrieval function returns items encoded in contexts most similar to the current context; as a result, it prefers to retrieve items that were encoded recently – since the state has drifted relatively little since recent items were encoded – and items that were encoded contiguously to items that were recently recalled.

This input-driven drifting context and similarity-based context retrieval and resetting bears a striking resemblance of EMRL's input-driven working memory state, similarity based retrieval, and reinstatement. This resemblance is especially resounding when querying is done with the cell state itself, rather than with a separate perceptual encoding. Increasing the resemblance is the fact that in practice the LSTM cell state displays high temporal autocorrelation, reminiscent of the input-driven *drift* that provides TCMs recency effect. EMRL's reinstatement, like TCMs resetting, provides a potential mechanism for the contiguity effect: reinstated cell states retrieved when recalling an item will increase the probability that the similarity-based retrieval mechanism will return the retrieved items temporal neighbors.

Future work may simulate EMRL in the list memory experiments that motivated TCM, using the supervised output vocabulary setup described in section 6.5, to test whether EMRL can reproduce the temporal and contiguity effects observed in humans and displayed by TCM. Of particular interest would be simulation experiments that train EMRL on a single task distribution and show that the trained agent reproduces the effects of interest in Howard and Kahana (2002) and the results of Hoskin et al. (2018) (see section 6.5 for details).

In sum, EMRL provides an alternative instantiation of some of the core ideas of TCM, and future work may pursue the possibility that EMRL can explain the empirical results that TCM accounts for.

7.0.2 Serial Recall

One of the most thoroughly studied topics in experimental psychology is memory for ordered lists of items (reviewed in Marshuetz, 2005). Botvinick and Plaut (2006) showed that a recurrent neural network model, set up as described in section 6.5, reproduces a variety of firmly established empirical effects pertaining to list length, primacy, recency, transposition, and fill-in errors. Botvinick and Plaut use these

results to argue that recurrent neural networks provide an effective model of working memory for serial order. Because EMRL is essentially a recurrent neural network working memory with extra learnable mechanisms, Botvinick and Plaut's simulations could all be run using EMRL. Because the immediate serial recall task does not yield much benefit to the use of episodic memory, it is likely that EMRL would simply learn to ignore the reinstated activations and behave like Botvinick and Plaut's recurrent networks. These observations suggest that EMRL provide be an effective model of serial recall. Future work may pursue this possibility, which, if borne out in simulation would position EMRL as a simple explanation covering a wide range of findings from list memory to decision making.

Of particular interest would be future work aimed at developing a single training distribution that would prepare an EMRL agent to reproduce human behavior in free recall tasks (see TCM section), serial recall tasks, and Hoskin et al.'s (2018) episodic reinstatement task.

7.0.3 Rational Accounts of Memory

EMRL is reminiscent of earlier rational accounts of memory, such as Anderson's account of the empirical availability of memories as adapted to the statistics of the environment Anderson and Schooler (1991), and Todd et al.'s (2009) model of working memory wherein the control mechanisms adapt to the environment. Because of its foundation in meta-learning, wherein the agent's behavior is determined by the properties of the training distribution, EMRL fits squarely with these theories. EMRL's LSTM could be considered an alternative implementation of the theory proposed by Todd et al., while the addition of a learned saliency value could bring about the frequency, recency, a prior exposure pattern effects described by Anderson and Schooler.

7.0.4 Gated Working Memory Models

A long thread of work has explored gated working memory models in cognitive neuroscience (Braver and Cohen, 2000; O'Reilly, Braver, Cohen, 1997; Frank et al., 2001; OReilly and Frank, 2006; Kriete et al., 2013; Chatham and Badre., 2015). This work develops the notion that control mechanisms are necessary around working memory, and the more recent work provides evidence that the basal ganglia plays a critical role in this gating process. This is all highly commensurate with the LSTM that comprises EMRL's working memory. The main difference is that the implementations used in simulations of these models carried more biological detail than EMRL. This increased detail provides an important proof-of-effectiveness for more biologically plausible implementations. EMRL complements this type of model by providing a more minimal sufficiency proof for the essential algorithmic/computational properties necessary to produce the behavior of interest.

7.0.5 Planning and Preplay

In addition to episodic memory, the hippocampus has been implicated in planning by "preplay" (see for review Ólafsdóttir et al., 2018), and a number of theories explain how this may work (of particular interest, Corneil and Gerstner, 2015; Mattar and Daw, 2018). EMRL's store of state representations provides the fodder for such a preplay algorithm. Future work may explore extensions of EMRL wherein a meta-learned planning algorithm – implemented e.g. by a recurrent network – queries the DND at each planning step in order to construct and evaluate plans. Such a system should in principle be able to take into account features of the environment observed only once when making plans, rather than relying only on features that have been incrementally learned by gradient updates over numerous exposures, as in past work on such recurrent planning agents (Weber et al., 2017). Further, future work may assess whether such a system might be suitable as a model of hippocampal preplay.

7.0.6 Replay, Consolidation, and CLS

One of the basic findings in the neuroscience of memory is the gradual reduction in reliance on the medial temporal lobe that memories undergo (Squire and Zola-Morgan 1991). This consolidation process is also associated with the extraction of statistical regularities in and relationships among many memories (e.g., McClelland et al. (1995), and has been linked with replay induced by the HC (Carr et al 2011, Deuker et al 2013). Future work may investigate extending EMRL with processes that operate over the DND contents to (1) tune the agent's policy or (2) improve the agent's environment model offline. From a machine learning perspective, such an algorithm would provide similar functionality to the replay buffers commonly used to train deep RL agents (Mnih et al., 2015), but would operate on high-level internal agent representations instead of environment state inputs. This would have the advantages of saving storage capacity, since the agent internal representations are typically smaller (although memory capacity is not typically a limiting factor), and it might be more suitable for more rapidly extracting high-level structure among memories. It would have the disadvantage that it would not be useful for training the perceptual functions that produce high level sensory and contribute to working memory representations. From a neuroscientific perspective, such an extension of EMRL could be useful for modeling the consolidation process as it relates to working memory, reward prediction errors, and episodic reinstatement. Future work will be necessary to assess whether this would provide a suitable model.

7.0.7 REMERGE and Flexible, Rapid, Relational Generalization

Prior work has demonstrated that humans and animals can behave appropriately in novel situations by computing relationships between items presented in separate past

episodes. Further, this ability does not rely on lengthy exposure to the individual items or long delay periods that would be necessary for consolidation (summarized in Kumaran and McClelland 2012). Kumaran and McClelland proposed an influential model that accomplishes this sort of generalization via recurrent processing of stored episodic memories. Future work may extend EMRL with recurrent functions - such as Santoro and colleagues' (2018) relational recurrent networks - that query its memory successively to compute relationships among stored memories, supplying the results of this computation to the agent's policy calculation. Such a system could be effective in tasks that require flexible, compositional, low-shot generalization.

7.0.8 Novelty Detection, Storage, and the Hippocampal-VTA Loop

Wimmer and Shohamy (2011) review prior work on interactions between the hippocampus and other learning systems in support of decision making. Perhaps the most prominent prior work[1]they discuss is Lisman and Grace's (2005 theory about novelty detection and learning rate modulation. In this theory, the hippocampus measures the novelty of the present state and conveys these novelty estimates to the VTA. Dopamine neurons originating in the VTA and terminating in the hippocampus fire in proportion to the signalled novelty. The dopamine released by these midbrain neurons in the hippocampus modulates the rate of Hebbian learning in hippocampal synapses. The effect of this hippocampal-VTA loop is to increase the strength of encoding in the presence of novelty. Lisman and Grace (2005) review a variety of evidence supporting this proposal.

Such a novelty detection loop suggests a solution to the problem of deciding when to save to EMRL's episodic memory. While the tasks simulated in this thesis afforded an obvious heuristic for deciding when to save – saving at the end of the episode – more realistic tasks will not have clear episode boundaries. Although the capacity of

storage systems like the DND is quite large, the brute-force strategy of storing a state for every experienced timestep will eventually run into capacity limitation. Perhaps more importantly, an agent that saves so much repetitious content may have more difficulty learning to retrieve memories because it is less likely to retrieve a useful memory by chance early in training.

EMRL could emulate the algorithm proposed by Lisman and Grace (2005) by using the distance of the query key from the nearest entry in the DND as a proxy for novelty of the current state. A distance threshold could be set for deciding whether or not to append the current working memory state to the DND, or alternatively, a scalar novelty indicator could be saved along with the working memory state. This novelty/salience scalar could decay over time (resetting to a high value if the memory was retrieved), and the associated memory trace could be deleted from the memory when the novelty scalar decayed below some threshold. Such heuristic solutions for deciding when to save/forget could be critical for EMRL's success in scaling to longer and more complex tasks.

7.0.9 Grid cells

The hippocampal formation is as famous for its role in state representation as it is for episodic memory. This observation leads to a number of intriguing questions for future research with EMRL: Why does the hippocampus serve both of these roles? There is probably a good reason for these two functions to be served by the same circuitry — how should this inform agent design? For example, do gridlike representations provide a favorable embedding space for memory lookup? Would

[1]Wimmer and Shohamy (2011) discuss one other hypothesized interaction between the hippocampus and another learning system, this time in support of of value transfer, i.e. sensory preconditioning. Sensory preconditioning refers to a learning effect wherein two stimuli are paired together, then one of the stimuli is paired with reward, then the animal associates the paired stimulus with reward even though that stimulus was never paired with reward. Wimmer and Shohamy review evidence that the hippocampus establishes a connection between two stimuli and the striatum establishes a connection between one of the stimuli and reward. Such a process might be captured by EMRL with the modifications proposed in the previous section on relational generalization.

minor modifications to EMRL induce place/grid cells in the lookup keys (see Banino et al., 2018 for related background)? Might such modifications provide performance benefits? Could integration of grid-like representations into the EMRL provide some insight about the relationship between these two roles of the hippocampal formation? Future work may investigate these questions.

7.0.10 Task Sets

Considerable prior work has focused on *task sets*, "active representations of behavioral strategies stored in long-term memory for driving action" (Collins and Koechlin, 2012). An intriguing connection exists between this well-developed concept and EMRL, which similarly saves representations of policies in the form of LSTM cell states in long-term memory and retrieves them to drive action. Much of the work on task sets has focused on the performance costs of switching between task sets (Sakai, 2008), and more recently, Collins and Koechlin (2012) and Donoso et al. (2014) introduced the notion that the PFC uses and monitors the efficacy of multiple task sets simultaneously, each of which can be modified in light of new experience. The model introduced by Collins and Koechlin executes Bayesian inference to determine which, if any, of the monitored strategies is effective in the current situation. If none is effective, the model creates and begins to tweak a new policy. Specifically, the model uses Dirichlet process mixtures, which have the capacity to add new strategies when appropriate.

Collins and Koechlin compared this model to human behavior in a task consisting of interleaved contextual bandits, where the mapping between context and rewarding arm changed intermittently and without notice. In one condition of the experiment, the mappings recurred, so that participants able to remember mappings discovered in the past were afforded an advantage. This task challenges participants to refine estimates of the mapping between contexts and arms while simultaneously gauging

whether or not the mapping has changed, necessitating a shift to a previously learned mapping or to the discovery of a new mapping.

EMRL is well situated to simulate this task, especially the version of EMRL discussed in Chapter 6 which queries with hidden states rather than separate perceptual cues. In this scenario, EMRL's working memory could be expected to learn to match stimuli to rewarding arms (a similar challenge to the bandit tasks of Chapter 3), while simultaneously monitoring for mapping changes. More presumptively, EMRL may also learn to query the DND when it detects mapping changes, looking for representations of state-action-outcome sequences similar to the recent sequence. Such history sequence-based lookup has been suggested previously by Plonsky et al. (2015), and this potential capability of EMRL is of interest in its own right. Plausibly, EMRL could learn to try actions appropriate to various mappings stored in the DND, continuing or rejecting them depending on their effectiveness. Finally, EMRL could learn to begin to refine a new policy if none of the policies stored in the DND prove effective. In short, EMRL has the architectural components necessary such that optimization on Collins and Koechlin's task might produce the policy arbitration behavior demonstrated by Collins and Koechlin. Analysis of EMRL's actions and query behavior after training could reveal whether the agents learns to approximate approximate rational inference. For instance, EMRL's behavior could be compared to that of the Dirichlet process mixture by measuring how well that model fits EMRL's behavior. This is left for future work.

It is worth pointing out a distinction between Collins and Koechlin's task sets and Harlow's learning sets, which in part motivated meta-RL (Harlow, 1949). The learning set in EMRL (and meta-RL) corresponds to the weights that encode exploration strategies, whereas the task set corresponds to the activations that encode behavioral strategies. In EMRL, the task set is defined by the learning set and the history of states, actions, and rewards in the current task. The learning set in contrast is

determined by optimization on all of the previously experienced tasks sampled from the task distribution. Future work reviewing literature on learning sets may reveal opportunities to leverage EMRL to understand the interplay between learning and task sets.

7.0.11 Extinction and Renewal

Extinction refers to the gradual reduction in a conditioned response when a previously reinforced stimulus is presented repeatedly without reinforcement. Renewal refers to the sudden return of the conditioned response to the stimulus after a change in context. The push to explain extinction and renewal has led to the development of RL theories which invoke state classification over adaptively expanding state spaces. Redish et al. (2007) proposed a model of ABA renewal, a specific form of renewal wherein a conditioned response is established in context A, extinction is carried out in context B, the renewal occurs when the animal is placed back in context A. In Redish and colleagues' model, temporal difference learning operates over an adaptively expanding set of states; each observation is compared to each of these prototypical states, and if one of the state prototypes matches, the observation is attributed to that state. If no state prototype matches the observation, a new state prototype is created.

This model predicts ABA renewal because the altered stimulus-reward contingencies in the context B cause the agent to infer a new state to which learning in that context is attributed. As a result, the agent's estimate of the association between the conditioned stimulus and reward in context A is unaffected by experience in context B. Thus, when the agent is returned to context A, it exhibits the conditioned response. Gershman et al. (2010) develop a model that is similar in spirit which performs Bayesian inference over an infinite capacity latent cause generative model.

Gershman and colleagues' model explains ABA as well as other forms of renewal, and casts state categorization as a process of rational inference.

EMRL captures some of the essential elements of these models: the nearest-neighbor lookup process constitutes a state classification mechanism, while its non-parametric storage provides an adaptively expanding catalogue of state representations along with learned contingencies associated with the states. The ABA renewal experiment could be simulated with EMRL by, for instance, presenting the agent with a stimulus and a pair of actions on every timestep along with a context identifier. One of the actions would correspond to the conditioned response, and would yield a relatively large amount of reward if taken in a context where the stimulus is rewarded, otherwise no reward. The other action would correspond to no-action, analogous to not making the conditioned response, and would always yield a relatively small amount of reward. This low reward represents the benefit to an animal of avoiding the effort of an unrewarded conditioned response.

If trained directly on such a task, EMRL would trivially learn to take the conditioned-response action in the first context (A), the other action in the second context (B), and the conditioned response action in the third context (A). Indeed, this would be a poor model of animal learning experiments because in those experiments, the animal is not exposed the ABA sequence many times to learn to exhibit extinction and renewal; instead, these behavioral features either are learned from previous experience or represent innate instincts. A more suitable, and potentially insightful way to model this task would be to train EMRL on a generic task distribution intended to capture the important abstract properties of the natural environment, such as the contingencies among contexts and stimulus-response associations. For example, a task distribution which samples multi-choice (bandit-like) tasks with random underlying MDPs where sometimes cue and/or context stimuli indicate the otherwise unobserved state of the MDP might induce inductive biases that

correspond to extinction and renewal. This sort of task distribution was discussed in Chapter 6 for modeling the experiments of Bornstein et al. (2017), Bornstein and Norman (2017), Wimmer et al. (2014), and Duncan and Shohamy (2016), and is inspired by work by (Plonsky et al., 2015). Fortuitously, the setup described in the previous paragraph matches the interface needed to model those studies: multiple possible actions, each with a latent reward probability; stimuli presented on each timestep; and context identifiers that remain fixed for multiple timesteps. This affords the possibility of simultaneously modeling these experiments and the extinction and renewal findings.

Approaching extinction/renewal using this type of environment model may reveal how these behavioral features arise from experience in an environment wherein stimulus-action-reward contingencies are modulated by context. Notably, the essence of this approach closely resembled that of Gershman et al. (2010), where the agent assumes a generative model in which stimulus-action-reward relationships are contingent on latent causes which are signalled by context. Future work may pursue such a modeling strategy, and may relate behavior of EMRL after training on a generic task distribution to the rational inference over Gershman et al.'s model.

7.0.12 Statistical Learning over Individuals and Events

In order to learn effectively about the categories of the natural world, humans must accurately parse their experience into persistent entities, i.e., individuals, and use those individuals as the basic unit of category learning. Marcus (2003) gives the example of a person living in an isolated environment with a dog who has three legs. While most of the dog encounters this person experiences will be with this particular dog, the person should identify that all of these experiences involve the same dog, and should not allow these experiences to lead to the belief that most dogs have three legs. Barsalou et al. (1998) present behavioral evidence that people in fact do track

individuals across events, and that they use representations of these individuals when learning to categorize.

This necessity to attribute multiple events to the same individual is closely related EMRL's reoccurrence of specific tasks. In the stochastic task process framework, the base distribution generates novel challenges with consistent structure, such as category relationships, which can be leveraged by the agent to learn quickly. The reoccurrence of specific tasks presents a challenge for an agent attempting to learn this structure: if the agent fails to attribute the multiple observations of the same task to a single persistent individual, it may learn a biased estimate of the base distribution, leading to suboptimal behavior in new problems. For example, an agent learning to solve multi-armed bandits, when tasked with the same identifiable bandit problem 90% of the time, should be able to avoid the belief that a large proportion of identifiably new bandit problems will have the same parameters as the reoccurring one. At the same time, it should be able to exploit the fact that it is familiar with the recurring problem.

It may be possible to use stochastic task processes to train agents like the one developed in this thesis to correctly infer latent individual structure and use that structure to estimate prior distributions over rewarding arms in novel and familiar contexts; in other words, to meta-learn to build representations of individuals in the LSTM hidden state like those discussed by Barsalou et al. (1998) and to use those representations for estimation in novel situations.

Bibliography

Thomas Akam, Rui Costa, and Peter Dayan. Simple plans or sophisticated habits? state, transition and learning interactions in the two-step task. *PLoS Comput Biol*, 11(12):e1004648, 2015.

John R Anderson and Lael J Schooler. Reflections of the environment in memory. *Psychological science*, 2(6):396–408, 1991.

JR Anderson. The adaptive character of thought. *Hillsdale, NJ: Earlbaum*, 1990.

Michael C Anderson, Jamie G Bunce, and Helen Barbas. Prefrontal–hippocampal pathways underlying inhibitory control over memory. *Neurobiology of learning and memory*, 134:145–161, 2016.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

David Badre and Anthony D Wagner. Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia*, 45(13):2883–2901, October 2007.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429, 2018.

Moshe Bar, Karim S Kassam, Avniel Singh Ghuman, Jasmine Boshyan, Annette M Schmid, Anders M Dale, Matti S Hämäläinen, Ksenija Marinkovic, Daniel L Schacter, Bruce R Rosen, et al. Top-down facilitation of visual recognition. *Proceedings of the national academy of sciences*, 103(2):449–454, 2006.

Helen C Barron, Raymond J Dolan, and Timothy EJ Behrens. Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature neuroscience*, 16(10):1492, 2013.

Lawrence W Barsalou, Janellen Huttenlocher, and Koen Lamberts. Basing categorization on individuals and events. *Cognitive Psychology*, 36(3):203–272, 1998.

Hannah M Bayer and Paul W Glimcher. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1):129–141, 2005.

Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. Deepmind lab. *arXiv preprint arXiv:1612.03801*, 2016.

Timothy E J Behrens, Mark W Woolrich, Mark E Walton, and Matthew F S Rushworth. Learning the value of information in an uncertain world. *Nat. Neurosci.*, 10 (9):1214–1221, September 2007.

Yoshua Bengio, Dong-Hyun Lee, Jorg Bornschein, Thomas Mesnard, and Zhouhan Lin. Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*, 2015.

David Blackwell and James B MacQueen. Ferguson distributions via pólya urn schemes. *The annals of statistics*, pages 353–355, 1973.

Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z Leibo, Jack Rae, Daan Wierstra, and Demis Hassabis. Model-free episodic control. *arXiv preprint arXiv:1606.04460*, 2016.

Aaron M Bornstein and Kenneth A Norman. Reinstated episodic context guides sampling-based decisions for reward. *Nature neuroscience*, 20(7):997, 2017.

Aaron M Bornstein, Mel W Khaw, Daphna Shohamy, and Nathaniel D Daw. Reminders of past choices bias decisions for reward in humans. *Nature Communications*, 8:15958, 2017.

Matthew Botvinick, Ari Weinstein, Alec Solway, and Andrew Barto. Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current Opinion in Behavioral Sciences*, 5:71–77, 2015.

Matthew M Botvinick and David C Plaut. Short-term memory for serial order: a recurrent neural network model. *Psychological review*, 113(2):201, 2006.

Jeffrey S Bowers and Colin J Davis. Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin*, 138(3):389, 2012.

Todd S Braver and Jonathan D Cohen. On the control of control: The role of dopamine in regulating prefrontal function and working memory. *Control of cognitive processes: Attention and performance XVIII*, pages 713–737, 2000.

Ethan S Bromberg-Martin, Masayuki Matsumoto, Simon Hong, and Okihide Hikosaka. A pallidus-habenula-dopamine pathway signals inferred stimulus values. *J. Neurophysiol.*, 104(2):1068–1076, August 2010.

C H Chatham and D Badre. Multiple gates on working memory. *Current Opinion in Behavioral Sciences*, 2015a.

Christopher H Chatham and David Badre. Multiple gates on working memory. *Current opinion in behavioral sciences*, 1:23–31, 2015b.

Christopher H Chatham, Michael J Frank, and David Badre. Corticostriatal output gating during selection from working memory. *Neuron*, 81(4):930–942, 2014.

Christopher Hughes Chatham and David Badre. Working memory management and predicted utility. *Frontiers in behavioral neuroscience*, 7:83, 2013.

G Chevalier and JM Deniau. Disinhibition as a basic process in the expression of striatal functions. *Trends in neurosciences*, 13(7):277–280, 1990.

Ignasi Clavera, Anusha Nagabandi, Ronald S Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt: Meta-learning for model-based control. *arXiv preprint arXiv:1803.11347*, 2018.

Jeremiah Y Cohen, Sebastian Haesler, Linh Vong, Bradford B Lowell, and Naoshige Uchida. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature*, 482(7383):85, 2012.

Jonathan D Cohen and Randall C O'Reilly. A preliminary theory of the interactions between prefrontal cortex and hippocampus that contribute to planning and prospective memory. *Prospective memory: Theory and applications*, pages 267–295, 1996.

Anne Collins and Etienne Koechlin. Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS biology*, 10(3):e1001293, 2012.

Dane S Corneil and Wulfram Gerstner. Attractor network dynamics enable preplay and rapid path planning in maze–like environments. In *Advances in Neural Information Processing Systems*, pages 1684–1692, 2015.

Erika Dahlin, Anna Stigsdotter Neely, Anne Larsson, Lars Bäckman, and Lars Nyberg. Transfer of learning after updating training mediated by the striatum. *Science*, 320(5882):1510–1512, 2008.

Nathaniel D Daw and Philippe N Tobler. Value learning through reinforcement: the basics of dopamine and reinforcement learning. In *Neuroeconomics (Second Edition)*, pages 283–298. Elsevier, 2014.

Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704, 2005.

Nathaniel D Daw, John P O'Doherty, Peter Dayan, Ben Seymour, and Raymond J Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095): 876–879, June 2006.

Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.

Ray J Dolan and Peter Dayan. Goals and habits in the brain. *Neuron*, 80(2):312–325, 2013.

Maël Donoso, Anne GE Collins, and Etienne Koechlin. Foundations of human reasoning in the prefrontal cortex. *Science*, 344(6191):1481–1486, 2014.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Katherine D Duncan and Daphna Shohamy. Memory states influence value-based decisions. *Journal of Experimental Psychology: General*, 145(11):1420, 2016.

Marcos Economides, Zeb Kurth-Nelson, Annika Lübbert, Marc Guitart-Masip, and Raymond J Dolan. Model-based reasoning in humans becomes automatic with training. *PLoS computational biology*, 11(9):e1004463, 2015.

Gilles O Einstein, Mark A McDaniel, Ruthann Thomas, Sara Mayfield, Hilary Shank, Nova Morrisette, and Jennifer Breneiser. Multiple processes in prospective memory retrieval: factors determining monitoring versus spontaneous retrieval. *Journal of Experimental Psychology: General*, 134(3):327, 2005.

C. Eliasmith. Attractor network. *Scholarpedia*, 2007.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

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.

Michael J Frank and Randall C O'reilly. A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. *Behavioral neuroscience*, 120(3):497, 2006.

Michael J Frank, Bryan Loughry, and Randall C O'Reilly. Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cognitive, Affective, & Behavioral Neuroscience*, 1(2):137–160, 2001.

S J Gershman and N D Daw. Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual review of psychology*, 68, 2017.

Samuel J Gershman, David M Blei, and Yael Niv. Context, learning, and extinction. *Psychological review*, 117(1):197, 2010.

Thomas Gisiger and Mounir Boukadoum. Mechanisms gating the flow of information in the cortex: what they might look like and what their uses may be. *Frontiers in computational neuroscience*, 5:1, 2011.

John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 148–177, 1979.

Jan Gläscher, Nathaniel Daw, Peter Dayan, and John 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.

Paul W Glimcher. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108(Supplement 3):15647–15654, 2011.

Mark A Gluck and Catherine E Myers. Psychobiological models of hippocampal function in learning and memory. *Annual review of psychology*, 48(1):481–514, 1997.

Daniel G Goldstein and Gerd Gigerenzer. Models of ecological rationality: the recognition heuristic. *Psychological review*, 109(1):75, 2002.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 2016.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 1735–1742. IEEE, 2006.

Todd A Hare, John O'Doherty, Colin F Camerer, Wolfram Schultz, and Antonio Rangel. Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of neuroscience*, 28(22): 5623–5630, 2008.

Harry F Harlow. The formation of learning sets. *Psychological review*, 56(1):51, 1949.

Kenneth D Harris. Stability of the fittest: organizing learning through retroaxonal signals. *Trends in neurosciences*, 31(3):130–136, 2008.

M. E. Hasselmo. Models of hippocampus. *Scholarpedia*, 2011.

Thomas E Hazy, Michael J Frank, and Randall C O'Reilly. Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1485):1601–1613, 2007.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8): 2554–2558, 1982.

A N Hoskin, A M Bornstein, K A Norman, and J D Cohen. Refresh my memory: Episodic memory reinstatements intrude on working memory maintenance. *bioRxiv*, 170720, 2017.

Abigail Novick Hoskin, Aaron M Bornstein, Kenneth A Norman, and Jonathan D Cohen. Refresh my memory: Episodic memory reinstatements intrude on working memory maintenance. *bioRxiv*, page 170720, 2018.

Marc W Howard and Michael J Kahana. Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(4):923, 1999.

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

Bernardo A Huberman, Peter LT Pirolli, James E Pitkow, and Rajan M Lukose. Strong regularities in world wide web surfing. *Science*, 280(5360):95–97, 1998.

Anthony Jang, Daniel Dillon, Michael J Frank, et al. Positive reward prediction errors strengthen incidental memory encoding. *bioRxiv*, page 327445, 2018.

Daphna Joel, Yael Niv, and Eytan Ruppin. Actor–critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw.*, 15(4):535–547, June 2002.

Lukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. *CoRR*, abs/1703.03129, 2017. URL `http://arxiv.org/abs/1703.03129`.

Steven W Kennerley, Mark E Walton, Timothy E J Behrens, Mark J Buckley, and Matthew F S Rushworth. Optimal decision making and the anterior cingulate cortex. *Nat. Neurosci.*, 9(7):940–947, July 2006.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

J N Kim and M N Shadlen. Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nat. Neurosci.*, 2(2):176–185, February 1999.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

Trenton Kriete, David C Noelle, Jonathan D Cohen, and Randall C O'Reilly. Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, page 201303547, 2013.

Dharshan Kumaran and James L McClelland. Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychological review*, 119(3):573, 2012.

B M Lake, R Salakhutdinov, and J B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350, 2015.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

Anders Lansner, Petter Marklund, Sverker Sikström, and Lars-Göran Nilsson. Reactivation in working memory: an attractor network model of free recall. *PLoS One*, 8(8):e73776, 2013.

Brian Lau and Paul W Glimcher. Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. Exp. Anal. Behav.*, 84(3):555–579, November 2005.

Sang Wan Lee, John P O'Doherty, and Shinsuke Shimojo. Neural computations mediating one-shot learning in the human brain. *PLoS biology*, 13(4):e1002137, 2015.

Mate Lengyel and Peter Dayan. Hippocampal contributions to control: The third way. In *Proc. of Neural Information Processing Systems, NIPS*, 2007.

Máté Lengyel and Peter Dayan. Hippocampal contributions to control: the third way. In *Advances in neural information processing systems*, pages 889–896, 2008.

Sarah Noami Lewis. The neural marketplace. 2014.

John E Lisman and Anthony A Grace. The hippocampal-vta loop: controlling the entry of information into long-term memory. *Neuron*, 46(5):703–713, 2005.

Kevin Lloyd, Nadine Becker, Matthew W Jones, and Rafal Bogacz. Learning to use working memory: a reinforcement learning gating model of rule acquisition in rats. *Front. Comput. Neurosci.*, 6:87, October 2012.

Adam H Marblestone, Greg Wayne, and Konrad P Kording. Toward an integration of deep learning and neuroscience. *Frontiers in computational neuroscience*, 10:94, 2016.

Gary F Marcus. *The algebraic mind: Integrating connectionism and cognitive science*. MIT press, 2003.

David Marr. Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 262(841):23–81, 1971.

David Marr. Vision: A computational investigation into the human representation and processing of visual information. mit press. *Cambridge, Massachusetts*, 1982.

Christy Marshuetz. Order information in working memory: an integrative review of evidence from brain and behavior. *Psychological bulletin*, 131(3):323, 2005.

Madoka Matsumoto, Kenji Matsumoto, Hiroshi Abe, and Keiji Tanaka. Medial prefrontal cell activity signaling prediction errors of action values. *Nat. Neurosci.*, 10 (5):647–656, May 2007.

Marcelo Gomes Mattar and Nathaniel D Daw. Prioritized memory access explains planning and hippocampal replay. *bioRxiv*, page 225664, 2018.

James L McClelland, Bruce L McNaughton, and Randall C O'reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

BL McNaughton. Associative pattern completion in hippocampal circuits: new evidence and new questions. *Brain Res Rev*, 16:193–220, 1991.

Frank A Middleton and Peter L Strick. Basal-ganglia 'projections' to the prefrontal cortex of the primate. *Cerebral Cortex*, 12(9):926–935, 2002.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

Kevin J Miller, Matthew M Botvinick, and Carlos D Brody. Dorsal hippocampus contributes to model-based planning. *Nature neuroscience*, 20(9):1269, 2017.

Jonathan W Mink. The basal ganglia: focused selection and inhibition of competing motor programs. *Progress in neurobiology*, 50(4):381–425, 1996.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Volodymyr Mnih, Adri Puigdomnech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous meth-

ods for deep reinforcement learning. In *Proc. of Int'l Conf. on Machine Learning, ICML*, 2016.

Hiroyuki Nakahara and Okihide Hikosaka. Learning to represent reward structure: a key to adapting to complex environments. *Neurosci. Res.*, 74(3-4):177–183, December 2012.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *CoRR, abs/1803.02999*, 2018.

Mike Oaksford and Nick Chater. *Rational models of cognition*. Oxford University Press Oxford, 1998.

H Freyja Ólafsdóttir, Daniel Bush, and Caswell Barry. The role of hippocampal replay in memory and planning. *Current Biology*, 28(1):R37–R50, 2018.

Randall C O'Reilly and Michael J Frank. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, 18(2):283–328, 2006.

Randall C O'reilly and James L McClelland. Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, 4(6):661–682, 1994.

Randall C O'Reilly and Yuko Munakata. *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. MIT press, 2000.

Randall C O'reilly and Jerry W Rudy. Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychological review*, 108 (2):311, 2001.

Camillo Padoa-Schioppa and John A Assad. Neurons in the orbitofrontal cortex encode economic value. *Nature*, 441(7090):223–226, May 2006.

Hilke Plassmann, John O'Doherty, and Antonio Rangel. Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *J. Neurosci.*, 27(37):9984–9988, September 2007.

Ori Plonsky, Kinneret Teodorescu, and Ido Erev. Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychological review*, 122(4): 621, 2015.

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adria Puigdomenech, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control. *arXiv preprint arXiv:1703.01988*, 2017.

D Purves, GJ Augustine, D Fitzpatrick, LC Katz, AS Lamantia, and JO McNamara. Neuroscience, 2nd edition. *Sunderland, MA: Sinauer Associates.[A systematic textbook of neuroscience.]*, 2001.

A David Redish, Steve Jensen, Adam Johnson, and Zeb Kurth-Nelson. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological review*, 114(3):784, 2007.

Samuel Ritter, Jane X Wang, Zeb Kurth-Nelson, and Matthew M Botvinick. Episodic control as meta-reinforcement learning. *bioRxiv*, page 360537, 2018a.

Samuel Ritter, Jane X Wang, Zeb Kurth-Nelson, Siddhant M Jayakumar, Charles Blundell, Razvan Pascanu, and Matthew Botvinick. Been there, done that: Meta-learning with episodic recall. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2018b.

Edmund Rolls. The mechanisms for pattern completion and pattern separation in the hippocampus. *Frontiers in systems neuroscience*, 7:74, 2013a.

Edmund Rolls. A quantitative theory of the functions of the hippocampal ca3 network in memory. *Frontiers in cellular neuroscience*, 7:98, 2013b.

Edmund T Rolls. Functions of the primate hippocampus in spatial and nonspatial memory. *Hippocampus*, 1(3):258–261, 1991.

Edmund T Rolls. An attractor network in the hippocampus: theory and neurophysiology. *Learning & Memory*, 14(11):714–731, 2007.

Edmund T Rolls. Attractor networks. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1):119–134, 2010.

Matthew F S Rushworth and Timothy E J Behrens. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat. Neurosci.*, 11(4):389–397, April 2008.

Brian F Sadacca, Joshua L Jones, and Geoffrey Schoenbaum. Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *Elife*, 5, March 2016.

Katsuyuki Sakai. Task set and prefrontal cortex. *Annu. Rev. Neurosci.*, 31:219–245, 2008.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016.

Adam Santoro, Ryan Faulkner, David Raposo, Jack Rae, Mike Chrzanowski, Theophane Weber, Daan Wierstra, Oriol Vinyals, Razvan Pascanu, and Timothy Lillicrap. Relational recurrent neural networks. *arXiv preprint arXiv:1806.01822*, 2018.

Juergen Schmidhuber, Jieyu Zhao, and MA Wiering. Simple principles of metalearning. *Technical report IDSIA*, 69:1–23, 1996.

Wolfram Schultz, Peter Dayan, and P Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

Hyojung Seo and Daeyeol Lee. Cortical mechanisms for reinforcement learning in competitive games. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 363(1511):3845–3857, December 2008.

Hyojung Seo, Dominic J Barraclough, and Daeyeol Lee. Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. *Cereb. Cortex*, 17 Suppl 1:i110–7, September 2007.

Moonsang Seo, Eunjeong Lee, and Bruno B Averbeck. Action selection and action value in frontal-striatal circuits. *Neuron*, 74(5):947–960, June 2012.

K Shima and J Tanji. Role for cingulate motor area cells in voluntary movement selection based on reward. *Science*, 282(5392):1335–1338, November 1998.

Daphna Shohamy and Nathaniel D Daw. Integrating memories to guide decisions. *Current Opinion in Behavioral Sciences*, 5:85–90, 2015.

David Silver. Ucl course on rl lecture 7, 2015. URL `http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching_files/pg.pdf`.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

R Angus Silver. Neuronal arithmetic. *Nature Reviews Neuroscience*, 11(7):474, 2010.

H Francis Song, Guangyu R Yang, and Xiao-Jing Wang. Reward-based training of recurrent neural networks for cognitive and value-based tasks. *Elife*, 6:e21492, 2017.

Bernhard P Staresina, Richard NA Henson, Nikolaus Kriegeskorte, and Arjen Alink. Episodic reinstatement in the medial temporal lobe. *Journal of Neuroscience*, 32 (50):18150–18156, 2012.

Clara Kwon Starkweather, Benedicte M Babayan, Naoshige Uchida, and Samuel J Gershman. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat. Neurosci.*, 20(4):581–589, April 2017.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In S A Solla, T K Leen, and K Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 1057–1063. MIT Press, 2000.

Yee Whye Teh. Dirichlet process. In *Encyclopedia of machine learning*, pages 280–287. Springer, 2011.

George R Terrell. *Mathematical statistics: A unified introduction*. Springer Science & Business Media, 2006.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

Edward Thorndike. *Animal intelligence: Experimental studies*. MacMillan, 1911.

Sebastian Thrun and Lorien Pratt. Learning to learn: Introduction and overview. In *Learning to learn*, pages 3–17. Springer, 1998.

Michael T Todd, Yael Niv, and Jonathan D Cohen. Learning to Use Working Memory in Partially Observable Environments through Dopaminergic Reinforcement. In D Koller, D Schuurmans, Y Bengio, and L Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1689–1696. Curran Associates, Inc., 2009.

Edward C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4): 189, 1948.

Alessandro Treves and Edmund T Rolls. Computational constraints suggest the need for two distinct input systems to the hippocampal ca3 network. *Hippocampus*, 2 (2):189–199, 1992.

Alessandro Treves and Edmund T Rolls. Computational analysis of the role of the hippocampus in memory. *Hippocampus*, 4(3):374–391, 1994.

Ken-Ichiro Tsutsui, Fabian Grabenhorst, Shunsuke Kobayashi, and Wolfram Schultz. A dynamic code for economic object valuation in prefrontal cortex neurons. *Nat. Commun.*, 7:12554, September 2016.

Oliver Vikbladh, Daphna Shohamy, and Nathaniel Daw. Episodic contributions to model-based reinforcement learning. In *Annual conference on cognitive computational neuroscience, CCN*, 2017.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.

Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. 2017. URL `arXivpreprintarXiv:1611.05763`.

Jane X Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6):860, 2018.

Xiao-Jing Wang. Synaptic reverberation underlying mnemonic persistent activity. *Trends in neurosciences*, 24(8):455–463, 2001.

XJ Wang. Attractor network models. In *Encyclopedia of neuroscience*. Elsevier Ltd, 2010.

Théophane Weber, Sébastien Racanière, David P Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adria Puigdomènech Badia, Oriol Vinyals, Nicolas Heess, Yujia Li, et al. Imagination-augmented agents for deep reinforcement learning. *arXiv preprint arXiv:1707.06203*, 2017.

Jeff Wickens. *A theory of the striatum*. Elsevier Science Inc., 1993.

Tom J Wills, Colin Lever, Francesca Cacucci, Neil Burgess, and John O'keefe. Attractor dynamics in the hippocampal representation of the local environment. *Science*, 308(5723):873–876, 2005.

G. Elliott Wimmer and Christian Buechel. Reactivation of reward-related patterns from single past episodes supports memory-based decision making. *Journal of Neuroscience*, 36(10):2868–2880, 2016.

G Elliott Wimmer and Daphna Shohamy. The striatum and beyond: contributions of the hippocampus to decision making. In *Decision Making, Affect, and Learning*. Oxford University Press, Oxford, 2011.

G. Elliott Wimmer, Erin Kendall Braun, Nathaniel D Daw, and Daphna Shohamy. Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. *Journal of Neuroscience*, 34(45):14901–14912, 2014.

Xiaoqian Xiao, Qi Dong, Jiahong Gao, Weiwei Men, Russell A Poldrack, and Gui Xue. Transformed Neural Pattern Reinstatement during Episodic Memory Retrieval. *J. Neurosci.*, 37(11):2986–2998, March 2017.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

David Zipser and Richard A Andersen. A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331 (6158):679, 1988.

Joel Zylberberg and Ben W Strowbridge. Mechanisms of persistent activity in cortical circuits: possible neural substrates for working memory. *Annual review of neuroscience*, 40:603–627, 2017.