

**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 copyright law. We do not charge any permission or usage fees for the publication of images of material in our 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 on the 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

Psychology

2023

Towards a Normative and Mechanistic Account of Cognitive Fatigue

Agrawal, Mayank

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp01wh246w44h>

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>

TOWARDS A NORMATIVE AND MECHANISTIC ACCOUNT
OF COGNITIVE FATIGUE

MAYANK AGRAWAL

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
PSYCHOLOGY
ADVISER: JONATHAN D. COHEN

SEPTEMBER 2023

© Copyright by Mayank Agrawal, 2023.

All Rights Reserved

Abstract

Cognitive fatigue and boredom are two phenomenological states that reflect overt task disengagement. In this dissertation, we present a rational analysis of the temporal structure of controlled behavior, which provides a formal account of these phenomena. We suggest that in controlling behavior, the brain faces competing behavioral and computational imperatives, and must balance them by tracking their opportunity costs over time. We use this analysis to flesh out previous suggestions that feelings associated with subjective effort, like cognitive fatigue and boredom, are the phenomenological counterparts of these opportunity cost measures, instead of reflecting the depletion of resources as has often been assumed. Specifically, we propose that both fatigue and boredom reflect the competing value of particular options that require foregoing immediate reward but can improve future performance: Fatigue reflects the value of offline computation (internal to the organism) to improve future decisions, while boredom signals the value of exploration (external in the world). We demonstrate that these accounts provide a mechanistically explicit and parsimonious account for a wide array of findings related to cognitive control, integrating and reimagining them under a single, formally rigorous framework. We follow this theoretical work with a behavioral and neuroimaging study that aims to explicitly test the mechanistic claim created by cognitive fatigue. The proposal of this new experimental paradigm is especially important with ‘ego depletion’ and its associated experimental paradigms falling out of favor. Chapter 2 introduces the theoretical framework and is separated into three parts, one for fatigue, one for boredom, and one for mind-wandering. The subsequent chapters evaluate the fatigue hypothesis. Chapter 3 introduces a behavioral study and its corresponding analyses. Chapter 4 discusses the neuroimaging portion, aiming to establish a link between replay and default mode activation. Chapter 5 concludes by placing this piece of work into a broader research agenda.

Acknowledgements

Wow, the last five years have been such a remarkable endeavor, and I have so many people to thank. First and foremost, I would like to thank Jon Cohen. Jon is truly one of the most selfless scientists I know. He is unbelievably generous with his time and energy, and his dedication to his students, his work, and the scientific community is unparalleled. Jon is incredibly adaptable and flexible to everyone's needs. He has always been there when I was floundering (which happens more often than would I like to admit), and he has always been patient with his wisdom. Jon champions the power of people more than anyone I know. Witnessing him realize these visions with his lab and the department has been an unforgettable and life-changing experience.

Next, I would like to thank Nathaniel Daw, who was somehow shepherded into this project (and thus my dissertation). Nathaniel is a scientist that champions integrity and precision, and this work would be nowhere near as strong if it wasn't for him. Nathaniel is unbelievably generous, too, willing to teach people anything from the basics of decision theory to how to code up complex neuroimaging analyses in multiple different software. I also think his Twitter is Princeton Neuroscience Institute's best marketing tool - and if you're reading this Nathaniel, I hope you are duly compensated (or at least, recognized) for your efforts.

Finally, I would like to thank Tom Griffiths, who wasn't a part of this dissertation but was an absolutely integral part of my PhD. I love the work we did together, and I think it has the potential to change how we do science long-term. Tom is one of the most nonjudgmental people I know, and his office (and Zoom room) was always an open and welcoming space for me when I needed to word-vomit. He would expertly extract the *gist* of my worries [78], tell me what I needed to hear, and would send me in a much more positive and happier direction.

I would also like to thank my other collaborators, specifically (1) those on this project: Sean Allan, Declan Campbell, and Marcelo Mattar; as well as (2) those on other projects: David Bourgin, Fred Callaway, Isaac Christian, David Danks, Rachit Dubey, Tyler Giallanza, Erin Grant, Josh Peterson, Daniel Reichman, and Simon Segert.

I would like to thank other professors and advisors that tremendously shaped my trajectory and often acted as a sounding board when I needed advice: David Danks, Sam Gershman, Adele Goldberg, Casey Lew-Williams, Tania Lombrozo, and Sue Sugarman. I would also like to thank my many undergraduate mentors and advisors that prepared me for a career in science: Geoffrey Aguirre, Lauren Kristine-Pryzant, Justin McManus, Jack Mostow, Tia Newhall, Pavan Ramkumar, Manuel Spitschan, and Dan Watkins.

To my cohort, Abia Alaoui-Soce, Xuechenzi Bai, Sori Baek, Fred Callaway, Chelsey Clark, Matt Hardy, Blake Laham, Emily Liquin, Mira Nencheva, we are awesome. There's not much more to say than that. We truly feel like a family. I love you all.

Blake and Matt, you two are my partners in crime. No amount of words can express my gratitude and love for you all. I can't wait to see how our shenanigans unravel in the future.

Lastly, my family. A few decades ago, a group of poor and rural villagers living in an oppressive caste-limited India got the golden ticket to come to the United States of America. They didn't have much but two things: (1) an incredible work ethic and (2) each other. And with those two things, they were able to build an incredible life for themselves and their children, letting people like me dream big and actually go for it. The Agrawals shaped every part of me, and I am and will forever be grateful for that upbringing.

To the Agrawal *parivaar*. I love y'all.

Contents

Abstract	3
Acknowledgements	4
1 Introduction	9
2 Theoretical Framework	11
2.1 Roadmap	12
2.2 Rational Models of Cognitive Control	13
2.3 Reinforcement Learning	14
2.3.1 Model-Free versus Model-Based Learning	15
2.3.2 Exploration vs Exploitation	16
2.4 Cognitive Fatigue	17
2.4.1 Empirical Findings	17
2.4.2 Hippocampal Replay	19
2.4.3 Discussion	25
2.5 Arousal and Boredom	28
2.5.1 Empirical Findings	29
2.5.2 Formalizing the Value of Information	31
2.5.3 Discussion	34
2.6 Mind-Wandering	36
2.6.1 Simulations	36
2.6.2 Discussion	38
3 Behavioral Study	39
3.1 Experimental Methods	40
3.1.1 Background: Two-Step Task	41

3.1.2	Task	41
3.1.3	Participants	43
3.1.4	Functional Localizer	44
3.2	Win-Stay, Lose-Shift (WSLS)	44
3.3	Hierarchical Bayesian Analyses	48
3.4	Discussion	50
4	Neuroimaging Analyses	52
4.1	Localizer Contrasts	53
4.2	Category Classifier	57
4.3	Task Contrasts	65
4.4	Default Mode Network	67
4.5	Discussion	70
5	Conclusion	71
5.1	Cost of Time	72
5.1.1	Big Data and Machine Learning for Theoretical Progress.	72
5.1.2	Integrating Theoretical and Practical Rationality in Philosophy	73
5.2	Concluding Remarks	73
A	Appendix	92
A.1	Theory Methods	92
A.1.1	Cognitive Fatigue	92
A.1.2	Boredom	93
A.1.3	Replaying to Explore	94
A.2	Aggregate Classifier Results	95

Chapter 1

Introduction

Questions surrounding qualia, instances of subjective experience, are a core part of the philosophy of mind. Yet, the questions are fundamentally psychological: *how* and *why* do people experience what they do? This dissertation aims to bring a highly interdisciplinary toolkit of cognitive psychology, computational modeling, analytical philosophy, and cognitive neuroscience to define and formalize the notion of mental effort. Specifically, we were interested in temporal dynamics, which can be captured by concepts such as flow, fatigue, and boredom. While all these concepts have been studied elsewhere, our work is likely the first to integrate them, demonstrating their similarities as well as their differences.

Our theoretical work, which was published in *Psychological Review* [5], is followed up by a behavioral and neuroimaging study [3] that aims to evaluate the novel mechanistic claim we propose for hippocampal replay. The theoretical work uses decision theory (specifically reinforcement learning) to taxonomize the phenomenological states of interest. We hope to reconcile at least a century's worth of work that has aimed to uncover the computational and neural bases of these states. The behavioral and neuroimaging study contributes by offering an experimental paradigm by which to precisely study the theoretical models' predictions. Phenomenology is an inherently difficult subject to (empirically) study, but the task of the psychologist is to do so nonetheless. With the previous experimental paradigm used to study ego depletion falling out of favor, a new design is needed that is reflective of the scientific advances we have made since. As the reader will see, the experiment was not as successful as hypothesized, but we believe nonetheless it serves as an excellent starting point by which to precisely study the elusive phenomena of fatigue and boredom.

Ultimately, the goal of this work is to ask more questions than it answers. Do we provide a meaningful decomposition of phenomenological states? Does our approach map onto both our intuition and our experimental paradigms? When it does not, is this a problem with our definition or our intuitive and experimental

language? Are these phenomena that can even be systematically studied? And, to what extent can we aim to leverage the language of computation to precisely denote them?

Chapter 2 introduces the theoretical framework and is separated into three parts, one for fatigue, one for boredom, and one for mind-wandering. The subsequent chapters evaluate the fatigue hypothesis. Chapter 3 introduces a behavioral study and its corresponding analyses. Chapter 4 discusses the neuroimaging portion, aiming to establish a link between replay and default mode activation. Chapter 5 concludes by placing this piece of work into a broader research agenda.

Chapter 2

Theoretical Framework

Learning¹ is one of the most widely studied processes in all of cognitive psychology. New tasks are often difficult, but they become easier – and subsequently, we perform them better – with practice [167, 183, 191, 8]. Computational models of learning propose that this is the result of minimizing prediction errors, and can be captured by connectionist models using backpropagation and/or reinforcement learning models using temporal difference learning [199, 38, 49].

However, to date, these models and most other formal theories of learning have largely failed to address the ubiquitously recognized subjective states of *cognitive fatigue* and *boredom*, and the changes in objective performance associated with these. Most theories predict that, with practice, there should be monotonic improvements in performance. In accordance with this prediction, greater practice does generally lead to progressive improvements in performance. For example, a participant training on a task for an hour every day will usually perform better after two weeks. Yet, most learning models do not take into account how the temporal characteristics of practice influence performance. They would naively predict that performance at the end of fourteen straight hours of practice is comparable to that at the end of the same amount of practice carried out periodically over two weeks². This is unlikely to be true [12, 91, 127, 208, 84, 123, 212, 81]. Specifically, after prolonged task engagement, it is all but certain that participants will feel *fatigued* or *bored*, and make considerably more errors (if they continue to perform for the full duration at all).

Fatigue has often been attributed to the consumption, and consequent diminution, of some resource (e.g., metabolic; [20, 21]), by analogy to the case of physical fatigue. In contrast, [115] proposed that exerting cognitive control (limitations of which may potentially characterize fatigue and boredom) may instead be

¹This chapter has been published. Agrawal M, Mattar MG, Cohen JD, Daw ND (2021). The Temporal Dynamics of Opportunity Costs: A Normative Account of Cognitive Fatigue and Boredom. *Psychological Review*.

²While the effects of massing vs. spacing of practice have been studied in several contexts, and several relevant factors have been identified [94, 28, 171, 56, 135], fatigue and/or boredom are almost certain to be dominant ones in this example.

accompanied by a sensation that signals opportunity costs that are a consequence of the limitation in the number of tasks that can be performed concurrently. That is, with the passing of time, it becomes increasingly possible that behaviors other than the one currently being performed offer opportunities for greater reward, and to which it would be more valuable to switch behavior [22, 89, 93]. However, this broad approach admits to many specific models. In addition to focusing on particular sensations (e.g., boredom), a fully specified opportunity cost model must satisfy two criteria: (1) it must define the nature of the alternative behaviors that give rise to the opportunity cost(s); and (2) it must account for the temporal dynamics of the phenomena it is meant to explain (e.g., the increase in fatigue and boredom over time). The goal of the present research is to specify such a formal theory for the cases of fatigue and boredom.

Here, we propose that one important class of opportunity costs arises from an intertemporal choice every agent must make: whether to sacrifice current reward in order to gather information that will result in greater reward later. Information gathering has value, which (if foregone, to instead pursue proximate reward) imposes an opportunity cost. As stated, this is the classic explore-exploit dilemma [77]. However, importantly, we extend this analysis to consider two different types of information-gathering actions. One corresponds to the standard treatment of explore-exploit: seeking out new opportunities in the external world, which can improve later choices. The second reflects an internal counterpart: offline processing by which one learns by thinking and mental simulation, again to compute decision variables that can improve future decisions. Both reflect a similar type of tradeoff between on-task performance and information gathering, but their values (and conversely, the opportunity costs for not pursuing them) have different dynamics with training. We identify the subjective feelings of boredom and fatigue with these specific, quantifiable decision variables which we propose our brain must compute and use for optimally allocating control.

2.1 Roadmap

First, we propose that fatigue adaptively signals the value of rest, and that the value of rest is derived from offline (internal) processing mechanisms such as hippocampal replay. We demonstrate how casting replay (a covert computational operation) as an instrumental action competing with overt behaviors leads to nontrivial dynamics of arbitration between replay and physical action in order to maximize future reward. Then, we propose that boredom tracks the value of exploring, here playing out via competition between different classes of overt action: information-seeking vs. exploitative. We offer a model that enables agents to navigate the explore-exploit tradeoff, expose its analogy to the case of replay, and demonstrate how the temporal dynamics of uncertainty lead agents to oscillate between different tasks. Finally, we integrate the two mechanisms of replay and exploration and examines new insights and problems arising from their interaction. Once viewed

in the same framework, the internal and external actions so far discussed separately can also interact, with consequences for their value, such as in the case of planning to explore. We examine some of these cases and speculate whether they may relate to additional subjective phenomena such as mind-wandering. In summary we propose that, rather than reflecting hindrances as is often assumed, fatigue and boredom reflect control optimizations that track the values of replay and exploration, respectively, and are used by agents to maximize long-term reward. A schematic of this decomposition is illustrated in Figure 2.1.

	Act	Replay
Exploit	Normal / Flow	Fatigue
Explore	Boredom	Mind-Wandering?

Figure 2.1: We paper partitions the constraints associated with the duration of cognitive control into two dimensions: action vs. replay and exploration vs. exploitation. First, we investigate fatigue, which we propose to be a signal for the value of replay while we propose boredom to be a bias towards exploration. We then integrate the mechanisms of replay and exploration; whether this corresponds to a currently studied phenomenology is unknown and is an important direction for future research.

Before we formally introduce our models of fatigue and boredom, we present background on cognitive control and reinforcement learning.

2.2 Rational Models of Cognitive Control

Cognitive fatigue has long been studied in psychology [201, 54], with one influential account, ‘ego depletion,’ suggesting that it reflects the consumption and subsequent diminution of a metabolic resource [20, 21]. Recently, however, this hypothesis has been called into question [115]; and multiple meta-analyses have challenged its empirical foundation, suggesting that the associated studies overestimated null effects

[33, 82, 170]. But beyond problems with the particular experiments to which it was applied, the metabolic hypothesis itself also seems mechanistically flawed: what exactly is this metabolic resource? Glucose has often been proposed, but there does not seem to be a relationship between executive function and glucose levels [134, 169, 76, 139, 181]. In fact, some of the largest glucose demands in the brain arise from visual processing [150], leading this account to predict, for instance, that face recognition should be more fatiguing than multi-digit arithmetic, though the opposite seems true in everyday life.

In contrast, normative models of cognitive control [190, 115, 119] propose that performance variation arises from the rational balancing of the costs versus benefits of different control strategies, rather than biological resource limitations. Although these accounts vary as to how they operationalize control (and thus what ultimately makes it costly or limited), the cost-benefit framing implies that performance decrements due to fatigue or boredom can be countered with incentives, shifting the tradeoff. Direct evidence for a rational basis of fatigue comes from a set of studies demonstrating that participants’ implicit theories about ‘ego depletion’ predicted the extent to which they exhibited depletion-like behavior [129, 36, 97, 98], suggesting that these individuals were conducting a metacognitive evaluation of the costs and benefits of rest and allocating control accordingly.

The current theory instantiates a rational model to explain fatigue and boredom. Agents’ actions in our model span two dimensions: physical versus mental and exploratory versus exploitative (Figure 2.1). Because an agent can only perform one action at a time (and in particular, because the mental actions we consider are assumed to exclude physical ones, for reasons later justified), each action (including covert, internal ones) comes with the opportunity cost of foregoing all other actions. The goal of the rational controller is to identify the sequence of actions that maximizes reward.

2.3 Reinforcement Learning

Reinforcement learning [199, 49] offers an integrative computational framework in which to implement a rational agent. Sequential decision problems in reinforcement learning settings are often modeled through a Markov decision process (MDP), a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$, in which $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\mathcal{R}(s)$ is the reward received in state s , $\mathcal{P}(s, a, s')$ is the probability of transitioning to from state s to state s' using action a , and γ is the discount factor. The policy $\pi : \mathcal{S} \mapsto \mathcal{A}$ determines with what probability an agent should perform action a when in state s and the agent’s goal in a reinforcement learning environment is to identify π^* , the policy that leads to maximum reward.

2.3.1 Model-Free versus Model-Based Learning

Two main classes of algorithms have emerged in reinforcement learning: model-free and model-based learning [199, 49]. These are exemplified by two different approaches for using trial-and-error experience to estimate the value of candidate actions so as to guide choices toward better options. Formally, we consider the *value function*, the expected cumulative future discounted reward $Q(s, a)$ for taking some action a in state s .

Model-free methods, such as Q-learning [213] estimate this function directly from experienced rewards over experienced state trajectories [141, 186, 185]. Here, an agent maintains an estimated function $Q(s, a)$, and updates it after every experienced state-action-reward-state transition (s, a, r, s') , according to the temporal difference learning backup rule:

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

in which α refers to the learning rate of the agent. In contrast, model-based learning [49, 197] learns an internal model of the environment (i.e. estimates the one-step dynamics $\mathcal{P}$ and rewards $\mathcal{R}$). Such a model can be used to *compute* $Q(s, a)$ by iterating steps and aggregating expected reward, formalizing a sort of mental simulation.

A main difference between these approaches is the computational work required to evaluate an action: model-based evaluation requires extensive internal iteration prior to action, whereas model-free values can be simply retrieved. Conversely, model-based evaluation is generally more accurate and flexible; this is because individual updates from Eq. 2.1 teach the agent about local rewards and costs, but working out their consequences for longer-run action-outcome relationships requires additional mental simulation (or many more experiential updates). Importantly, such mental simulation can “teach” the agent how to make better choices in future, but without actually collecting new information from the environment — instead, by discovering the consequences of information already known. For these reasons this computational distinction has been employed in neuroscience as a model of the tradeoffs between thinking and acting [49, 105]. The general idea is that the brain can either act immediately according to (fast, potentially inaccurate) model-free values, or spend time computing better (more accurate) model-based ones. This leads to a speed-accuracy tradeoff and a rational account of many phenomena of habits, automaticity, compulsion, and slips of action [49, 105, 157, 158, 117, 106, 110, 111, 189].

The current theory employs a finer-grained version of this idea, based on the Dyna framework [198], which learns values from experience using Eq. 2.1, but also makes decisions about whether to improve these using individual steps of model evaluation and, if so, which ones [130]. The core tradeoff — whether to act,

or delay action to produce more accurate evaluations — and the logic of its cost-benefit resolution remain the same. Mattar & Daw [130] proposed that the brain implements the steps of model evaluation by replaying trajectories in hippocampus. This theory will be the basis of our first analysis, where we suggest fatigue tracks the value of such replay.

2.3.2 Exploration vs Exploitation

Reinforcement learning also offers an analysis of the explore-exploit tradeoff. When picking which restaurant to visit, whether to date a potential partner, or what research programs to pursue, humans must decide whether to *exploit* options they know are rewarding or forego those to *explore* new options that may potentially be even more rewarding. The explore-exploit tradeoff has long been studied in computer science and has recently attracted increasing attention in psychology and neuroscience [101, 199, 50, 39, 218, 133, 72, 187].

The value of exploratory actions, in principle, is that the agent may learn something from them that improves their future choices — and thus their future earnings. The classic decision-theoretic analysis of the explore-exploit dilemma in problems such as bandit tasks [77] attempts to quantify this long-run value directly by computing actions' expected future value taking into account the possible effects of learning on later choices and rewards. This gives rise to a difference, for each action, between its nominal value Q based on current knowledge, versus its expected long-run value including the improvement due to learning. For our purposes, we can generically express this as the sum of a baseline value Q and an additional increment, called value of information (VOI):

$$Q_{VOI}(s, a) = Q(s, a) + VOI(s, a) \quad (2.2)$$

In practice, the VOI depends on the task: for instance, how it is broken up into repeated episodes, and what is shared between them (and/or other tasks) that can be learned to improve later performance. In general, although it can be defined formally, computing it exactly is typically intractable except for particular special cases. However, there are many heuristics and approximations to it. These can be added to nominal value Q to help the agent pursue exploratory behavior over immediate reward in circumstances in which that leads to greater overall (i.e., long-term) reward. A typical example is the Upper Confidence Bound (UCB) algorithm [16], which proposes the VOI in a multi-armed bandit setting to be

$$VOI = \sqrt{\frac{2 \ln n}{n_i}} \quad (2.3)$$

in which n is the total number of trials and n_i is the number of trials arm i has been selected. According to

this formula, VOI increases as time passes, and decreases with the number of times a given action is selected. Like many such heuristics, this quantity is a rough proxy for uncertainty about the value Q , which in turn measures how much can be learned about the task. This framework will be the basis of our analyses in which we suggest increasing boredom reflects a bias towards exploration.

2.4 Cognitive Fatigue

A nearly ubiquitous observation is that, as we exert mental effort, we experience fatigue and eventually want to take a break. Furthermore, after taking a break, we may feel rejuvenated and willing to perform the task again. Fatigue has long been associated with rest [109, 145], with Edward Thorndike defining fatigue as “that diminution in efficiency *which rest can cure* (emphasis ours)” over one hundred years ago [202]. Here, we propose the value of rest is derived from offline computational processes such as hippocampal replay. But first, we summarize the evidence any rational theory of fatigue must explain.

2.4.1 Empirical Findings

We identify three canonical effects in the literature a rational model of fatigue needs to explain: (1) why rest is valuable; (2) when an agent should switch between rest and action; and (3) why difficult tasks are more fatiguing than easier tasks.

Rest Helps Performance

Several studies have examined the effects of rest in mitigating decrements in performance with time on task [23, 172, 85, 86]. For example, [85] demonstrated the benefit of rest by having participants carry out a vigilance task [127] that was interrupted by either a rest period or another task before resuming the initial task. In their first experiment, they found that participants who were given a rest period performed better post-interruption than those who remained on task. In a second experiment, [85] expanded the set of interruption conditions from two to five (rest, continuation, verbal match to sample, letter detection, or spatial memory). They found that the restorative effects of the interruption were predicted by the degree to which it involved a task that was distinct from (i.e., did not overlap) with the vigilance task: those in the rest condition performed the best post-interruption, and those in the continuation condition performed the worst. Furthermore, participants in the verbal match-to-sample condition, which had the least amount of overlap with the vigilance task, performed the second best, and participants in either the letter detection or spatial memory conditions (which had partial overlap with the vigilance task) performed better than those in the continuation condition

but worse than those in the verbal match to sample condition. Thus, the more that the interruption involved a task similar to the vigilance task, the less it helped.

Arbitration Between Labor and Leisure

Kool & Botvinick [109] proposed that the choice of how much and for how long to engage in a cognitively-demanding task vs. rest reflects a valuation of mental effort and rest as non-substitutable “goods” (i.e., forms of utility), that can be described using the same approach used to analyze the labor/leisure tradeoff in economics [151]. To demonstrate this, they conducted an experiment in which participants were allowed to alternate as they wished between doing three-back and one-back (the latter of which effectively played the role of ‘rest’) versions of the N-back task [107] for one hour, with increased time on the three-back resulting in increased compensation. They observed that participants sought a balance between doing the three-back and one-back, which they described in terms of a joint concave utility function combining labor and leisure. Evidence for this concavity came from experiments manipulating the fixed and variable wages, and measuring the direction in which the tradeoff changed. This provided a formally rigorous description of the tradeoff between mental effort and rest, in which fatigue can be interpreted as reflecting the value of rest and, as suggested by the authors, a normative framework for relating the tradeoff to rational models of control allocation [190]. However, this account did not provide an explanation for the value of rest.

Difficult Tasks are More Fatiguing

Many fatigue studies that have reported depletion-like effects follow a sequential-task format: Engage the participant in a first task that is ‘depleting,’ and demonstrate there is a negative effect on performing a subsequent second task. For example, [25] conducted a study over the span of six hours. Participants performed either the ‘easy’ tasks of a one-back and one-switch³ or the ‘hard’ tasks of a three-back and twelve-switch. Every thirty minutes, participants were given a block of intertemporal choice trials. The depletion effect was measured by the amount of discounting in these trials. Although performance on the primary tasks (N-back and N-switch) was comparable across groups and consistent throughout the experiment, participants in the ‘hard’ condition made increasingly more impulsive choices (i.e. discounted more heavily) over the course of the experiment, whereas those in the ‘easy’ condition did not show this effect. Assuming that increased impulsivity reflects fatigue, the results from this experiment suggest that participants in the ‘hard’ condition were more fatigued than those in the ‘easy’ condition.

³In an N-switch block, the participant switches between two tasks N times.

Discussion

Rest thus seems to play a vital role in understanding the normative basis of fatigue. Whereas rest is sometimes assumed to reflect the lack of activity, an extensive body of evidence in the memory literature now suggests that it is a state in which the brain engages in offline processing mechanisms such as planning and consolidation [131, 200, 32, 156, 210]. This suggests a grounding for the benefits of wakeful rest — i.e., the value of planning and consolidation, or more particularly the improvement in future reward those processes may achieve. If so, the agent should induce a state of ‘rest’ when its estimated value surpasses the estimated value of physical action⁴. We propose that this value is represented by the phenomenological experience of fatigue. Below, we discuss hippocampal replay as one mechanism of offline processing that has a quantifiable value.

2.4.2 Hippocampal Replay

Neurons in hippocampus called “place cells” are famously tuned to spatial locations, that is they tend to respond when the organism is in a certain location [155, 144]. Interestingly, they also fire in coordinated patterns that appear to represent trajectories removed from the animal’s location. Hippocampal replay refers to the physiological phenomenon in which hippocampal place cells fire in sequential patterns during periods of sleep and awake rest [64, 53, 47, 104, 80]. Replay events are commonly observed during epochs of high-frequency oscillatory activity in the hippocampus known as ‘sharp wave ripples.’. When compared with the spatial locations represented by the place cells, the replayed sequential patterns often correspond to spatial trajectories – both experienced and novel – in the animal’s physical environment [217, 147, 124, 116]. Though hippocampal replay has been most frequently observed and characterized in rodents, recent studies have also begun to characterize a corresponding phenomenon in humans during periods of rest [74, 178, 140, 121, 220, 184, 60, 122].

Importantly, sharp wave ripples – and associated replay – occur one trajectory at a time, when an animal is standing still, resting, or asleep. During active locomotion, hippocampus predominantly represents the animal’s current location (or oscillates a bit ahead and behind it, in sync with a distinct mode of theta-band oscillation in the EEG). This is important in the current context, because it means that hippocampal replay events carry an opportunity cost: they are exclusive of active locomotion. It is likely that this reflects contention for a shared resource: the hippocampal representation of location, which can only represent one location at a time, and thus can’t be used simultaneously to represent physical presence at one location but replay of another.

⁴A similar argument has been made for sleep, suggesting that sleep is the ‘price that the brain pays for plasticity’ [203, 204, 205]

Mattar and Daw (2018)

Mattar & Daw [130] (henceforth referred to as M&D) investigated the utility of hippocampal replay within a reinforcement learning setting. They proposed that replay acts as the physiological instantiation of a step of model-based value computation over that location [199, 49]. Under this model, replay has the potential to affect the agent’s future behavior, and therefore the potential to increase its expected future reward. The place cells activated during replay events are assumed to correspond to the experiential states the agent is simulating.

Replay has been proposed as a mechanism by which the model-based system can be used to accelerate learning relative to traditional model-free algorithms. While the latter, such as Q-learning, have been proven to converge to the optimal policy after sufficient experience [213], this process can be slow in practice because it relies on interactions with the external world. Replay can be thought of as a mechanism by which simulated experience using the model-based system is substituted for physical experience [198]. This is useful, in turn, because experience is actually playing two roles in an algorithm like Q-learning. It is both interacting with the world to gather information about how a task works, e.g. the location of rewards, but also propagating that information along experienced trajectories to work out its consequences for distal actions. The latter function (though not the former) can also be accomplished by mental simulation. For instance, even once you know the rules of chess completely, it takes further computation to elaborate their consequences for the best moves in particular situations. If this simulated experience is faster than physical experience and/or selected through a priority metric [162, 143], the agent can converge to the optimal policy quicker, and thus increase future reward, as opposed to relying exclusively on physical experience.

M&D derived the value of a single replay event, called the Expected Value of Backup (EVB), and ran a set of simulations under the assumption that agents replay the state-action pair (s_k, a_k) with the highest EVB at the beginning and end of a trial.

$$EVB(s, s_k, a_k) = \mathbb{E}_{\pi_{\text{new}}} \left[\sum_{i=0}^{\infty} \gamma^i R_{t+i} \middle| S_t = s \right] - \mathbb{E}_{\pi_{\text{old}}} \left[\sum_{i=0}^{\infty} \gamma^i R_{t+i} \middle| S_t = s \right] \quad (2.4)$$

$$= Gain(s_k, a_k) \times Need(s, s_k) \quad (2.5)$$

The *Gain* corresponds to the expected increase in expected reward following a visit to the replayed state (since this is the only state in which choice can be affected by a one-step backup) and can be expressed as:

$$Gain(s_k, a_k) = \sum_{a \in A} Q_{\pi_{\text{new}}}(s_k, a) \pi_{\text{new}}(a | s_k) - \sum_{a \in A} Q_{\pi_{\text{old}}}(s_k, a) \pi_{\text{old}}(a | s_k) \quad (2.6)$$

That is, the change in expected future reward Q expected following a visit to state s_k , due to following the new policy π_{new} resulting from the computation, vs. following the status quo policy π_{old} . The *Need* term corresponds to the expected number of (delay discounted) future visits to that state:

$$\text{Need}(s, s_k) = \mathbb{E} \left[\sum_{i=0}^{\infty} \gamma^i \delta_{s_{t+i}, s_k} \mid s_t = s \right] = M(s, s_k) \quad (2.7)$$

Here, δ is the Kronecker delta function, so the Need is the expected future discounted occupancy for the contemplated state s_k starting in the current state s . This in turn can be obtained from the successor representation M [51, 75], estimated as $\widehat{M}$, for the state pair (s, s_k) .

M&D showed that this model accounts for a wide range of empirical findings in the replay literature, including in particular the reported predominance of forward and reverse replay in the beginning and end of a trial, respectively.

Expected Value of Backup with Cost (EVB_C).

While the M&D model provides a rationale for *which* experiences an agent should replay [162, 143, 180], it does not directly address the question of *when* an agent should replay. Thus, we extend the original M&D model to provide a normative answer this question, by taking into account not only the benefits that replay has for performance, but also the opportunity cost that it carries in time; that is, by delaying the opportunity for reward.

Formally, $\text{EVB}_C(s, s_k, a_k)$ is the expected increase in reward resulting from replaying the state-action pair (s_k, a_k) while in state s and executing the corresponding Bellman backup (i.e. temporal difference learning update, as in Equation 2.1 but for a simulated rather than experienced step), minus the amount of reward lost due to the time it takes to replay that state-action pair.

$$\text{EVB}_C(s, s_k, a_k) = \mathbb{E}_{\pi_{\text{new}}} \left[\sum_{i=\tau}^{\infty} \gamma^i R_{t+i} \mid S_t = s \right] - \mathbb{E}_{\pi_{\text{old}}} \left[\sum_{i=0}^{\infty} \gamma^i R_{t+i} \mid S_t = s \right] \quad (2.8)$$

$$= \gamma^\tau v_{\pi_{\text{new}}}(s) - v_{\pi_{\text{old}}}(s) \quad (2.9)$$

$$= \gamma^\tau v_{\pi_{\text{new}}}(s) - \gamma^\tau v_{\pi_{\text{old}}}(s) + \gamma^\tau v_{\pi_{\text{old}}}(s) - v_{\pi_{\text{old}}}(s) \quad (2.10)$$

$$= \gamma^\tau (v_{\pi_{\text{new}}}(s) - v_{\pi_{\text{old}}}(s)) + \gamma^\tau v_{\pi_{\text{old}}}(s) - v_{\pi_{\text{old}}}(s) \quad (2.11)$$

$$= \gamma^\tau \text{EVB}(s, s_k, a_k) + \gamma^\tau v_{\pi_{\text{old}}}(s) - v_{\pi_{\text{old}}}(s) \quad (2.12)$$

$$= \gamma^\tau \text{EVB}(s, s_k, a_k) - (1 - \gamma^\tau) v_{\pi_{\text{old}}}(s) \quad (2.13)$$

$$= \gamma^\tau \text{EVB}(s, s_k, a_k) - (1 - \gamma^\tau) \sum_a \pi_{\text{old}}(a|s) q_{\pi_{\text{old}}}(s, a) \quad (2.14)$$

Here, the first term corresponds to the Expected Value of Backup (EVB) discounted by τ which is the ratio of time it takes to replay versus act⁵. This discounting indicates that the benefits of replay can only be accrued after the time required to replay, τ , has elapsed. The second term is the reward lost due to the time it takes to replay: $\sum_{a \in A} Q_{\pi_{\text{old}}}(s, a) \pi_{\text{old}}(a|s)$ is the expected discounted future reward that would be available if the agent started acting immediately, and $\gamma^\tau \sum_{a \in A} Q_{\pi_{\text{old}}}(s, a) \pi_{\text{old}}(a|s)$ is the same quantity adjusted by the passage of τ . Their subtraction – i.e. $(1 - \gamma^\tau) \sum_{a \in A} Q_{\pi_{\text{old}}}(s, a) \pi_{\text{old}}(a|s)$ – is thus the reward forgone due to the time it takes to replay.

Hippocampal Replay as the Value of Leisure

The M&D model, and our subsequent EVB_C extension, provides a quantitative value to ‘rest’ if the agent is engaging in replay during these rest states. Defining EVB_C^* as $\max EVB_C(s, \cdot, \cdot)$, a rational agent should replay the most valuable location, $\arg \max EVB_C(s, \cdot, \cdot)$, as long as its value $EVB_C^* > 0$. Hence, if EVB_C^* is positive, replaying is more valuable than acting, a situation which we propose is subjectively sensed as fatigue. If EVB_C^* is negative acting is more valuable than replaying, and thus the agent should physically act instead of resting. Thus, the agent is optimizing the intertemporal tradeoff between acting (providing a more immediate opportunity for reward) and replaying (providing an opportunity for greater but later reward). This insight may help to rationalize the labor and leisure tradeoff that has been described for cognitive control [109, 152, 153, 93, 57].

Results

Figure 2.2 plots the replay behavior of an agent pursuing a specified reward in a gridworld, using the EVB_C -driven replay algorithm described above, for different values of τ (all details of simulations are in the Methods section of the Appendix). Three phases of replay behavior can be seen in these plots. In the first, the agent does not replay because there is no knowledge of the reward structure in the first trial, and thus there is no value to replay. Instead, the agent is accumulating experience to build an internal model of the environment (specifically, it is discovering rewards and developing its successor representation). In the second phase, the agent replays extensively because it has a good internal model but has still not fully developed and refined its value function, and thus can still gain by adjusting its Q-values through replay as well as action. Due to the expanded scope of replay (i.e. the ability to replay *any* experience), the speed benefit of replay, as well as the low value of action, EVB_C is positive in this regime. This is the phase we identify with fatigue. Lastly, in the third phase, the agent stops replaying because its value function has become sufficiently good that the

⁵One estimate of τ is 0.04, that comes from the speed of sharp wave ripples in the hippocampus; these occur at approximately 1,000 cm/s [165] as compared to the speed of running on a track, which is approximately 40 cm/s [214]. In humans, a recent study by wimmer2019episodic suggests that replay can be sixty times faster than physical action.

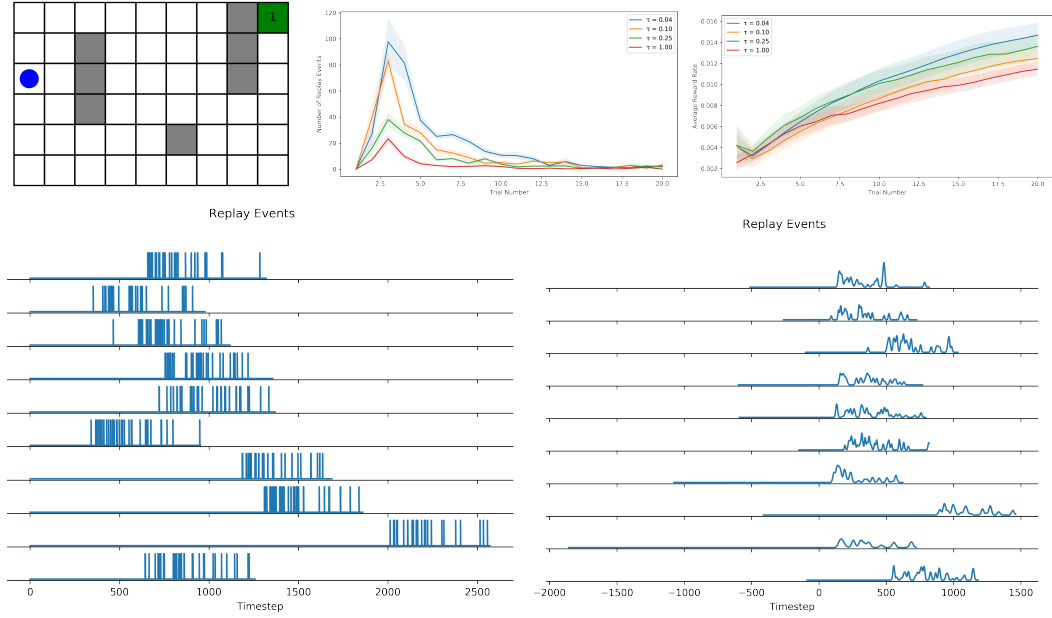


Figure 2.2: Simulation results for gridworld agent. (Top Left) Gridworld environment used for simulations. (Top Middle) Number of replay events over the course of multiple trials for different values of τ . (Top Right) Average reward rate for different values of τ . All error bars indicate ± 1 SEM. (Bottom Left) Spikes indicate individual replay events. (Bottom Right) A smoothed version of the left panel. Each trial is also shifted by the amount of time it took for the first trial (which is purely random exploration).

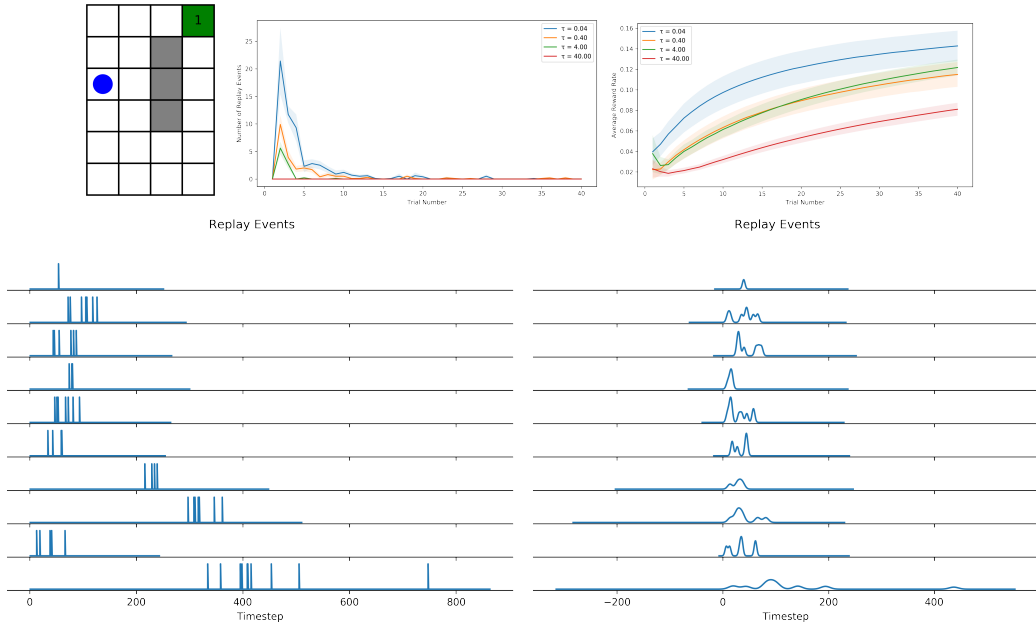


Figure 2.3: Simulation results for running easy gridworld agent. (Top Left) Easy gridworld environment used for simulations. (Top Middle) Number of replay events over the course of multiple trials for different values of τ . (Top Right) Average reward rate for different values of τ . All error bars indicate ± 1 SEM. (Bottom Left) Spikes indicate individual replay events. (Bottom Right) A smoothed version of the left panel. Each trial is also shifted by the amount of time it took for the first trial (which is purely random exploration).

opportunity costs of replay exceed its benefits [209, 207]. This third regime — in which the value function has converged, $EV B_C$ is negative, and behavior is executed without further deliberation — can be thought of as a transition to fully model-free, automatic processing. According to our model, cognitive fatigue, and corresponding periods of rest, arise during the preceding controlled, deliberative phase.

Our model explains the three canonical effects outlined earlier. The restorative power of rest demonstrated in [85] can be explained in a straightforward way in terms of replay: rest provided the participants an opportunity for replay that facilitated learning and later performance (and also diminished the need for subsequent replay, which itself would compete with task performance). The second experiment, on the effects of filling a task interruption phase with different interfering tasks, can be explained in the same terms, if it is assumed that the opportunity for replay during the interruption was (inversely) related to the extent to which the task performed during the interruption shared processing resources with those engaged by the vigilance task (e.g., visual encoding and identification of letters). There is strong evidence in the literature that tasks that share processing resources, and risk interference with one another as a consequence, rely on control to mitigate such interference by ensuring that only one is performed at a time [149, 136, 174, 146]. Assuming the same holds for replay (i.e., that it relies on the same perceptual and decision making mechanisms engaged by overt performance), then the more the interruption task shared resources with the vigilance task, the less opportunity it provided for replay of the vigilance task and its salubrious effects.

Figure 2.2 also grounds the labor-leisure tradeoff of `kool2014labor` in normative terms, if it is assumed that leisure corresponds to time allocated for replay. Rather than positing leisure as intrinsically valuable, we demonstrate how the temporal dynamics of its value as an opportunity for replay (mathematically derived in our model) leads it to be either greater than or less than the value of action at different points in time. Accordingly, a rational agent should arbitrate between periods of action and replay, based on which maximizes future reward.

Lastly, to model the relationship between task difficulty and fatigue, we evaluate our agent on an easier gridworld, shown in Figure 2.3. Since the agent takes less time to develop automaticity in this task (i.e., learn the relevant value function), the overall replay behavior (and hence fatigue) is reduced. Thus, our model is able to demonstrate the relationship between fatigue and task difficulty. Specifically, more difficult tasks are more fatiguing because replay has greater value relative to immediate action in tasks in which fully learning the value function is slower. Therefore, the three-back is fatiguing, as demonstrated in [25]. Conversely, once the value function is learned and the task can be executed automatically without further replay, there is little value in offline processing and thus the model predicts less or no fatigue, as demonstrated in the one-back condition of [25].

2.4.3 Discussion

The model we propose suggests that the role of leisure goes beyond what it is commonly thought to be “doing nothing.” A large body of evidence suggests that, during states of rest, agents replay past memories to help improve future performance. Rational agents should thus induce these states when they are valued higher than action. We propose that this explains the phenomenological experience, and corresponding behavioral observations, of cognitive fatigue. Furthermore, the need for arbitration between replay and action – as well as the competition between replay and any intervening tasks (such as in the study of [85] above) – can be explained as a result of the inability to simultaneously use the same processing resources for different purposes at the same time. Since the purpose of replay is to improve the representations used for action, use of these to replay one set of stimulus-actions sequences while physically engaging in another would produce conflict, and thus both cannot be done concurrently. This is consistent with most hippocampal replay studies to date, which show that sharp wave ripples are rarely observed during locomotion.

Benefits and Limitations of Replay

Updating learned action values is one benefit of hippocampal replay, but it is plausible (and probable) that there are other benefits. The complementary learning systems framework [131, 113, 179] suggests that another benefit of offline, hippocampal replay is preventing catastrophic interference that can occur in gradient learning due to the high autocorrelation of online experience [138]. Similarly, understanding the extent to which replay during awake rest differs from that during sleep will help inform our understanding of the benefits of the different offline processing mechanisms. There may also be some limitations of replay. [46] proposed that mental simulation may reflect a noisy form of physical simulation. Thus, physical action and experiential learning may be more valuable in situations in which it is difficult to build a model of the environment. However, mental simulation may be more useful (relative to direct trial-and-error learning) for discovering delayed action-outcome relationships in multi-step sequential tasks, such as spatial tasks, social situations, and games. Additional research characterizing different offline processing mechanisms according to these factors will be valuable in generating a more precise understanding of how agents should rationally arbitrate between action and rest states.

Intra-Trial Dynamics

Fatigue studies generally consider the number of trials or the time on task as the causally relevant measure. The model proposed here suggests that learning is a mediating variable, and offers a more temporally fine-grained analysis, making quantitative predictions in terms of the states in a Markov decision process.

Consistent with experimental observations, some points are better than others for replay within individual trials; replay during rodent navigation tasks most often occurs at the start and end of trials as well as at choice points [32, 156]. The dynamics of the $EVBC$ agent are shown in Figure 2.4, and thus we predict that sequential tasks should have specific patterns of fatigue dynamics within individual trials.

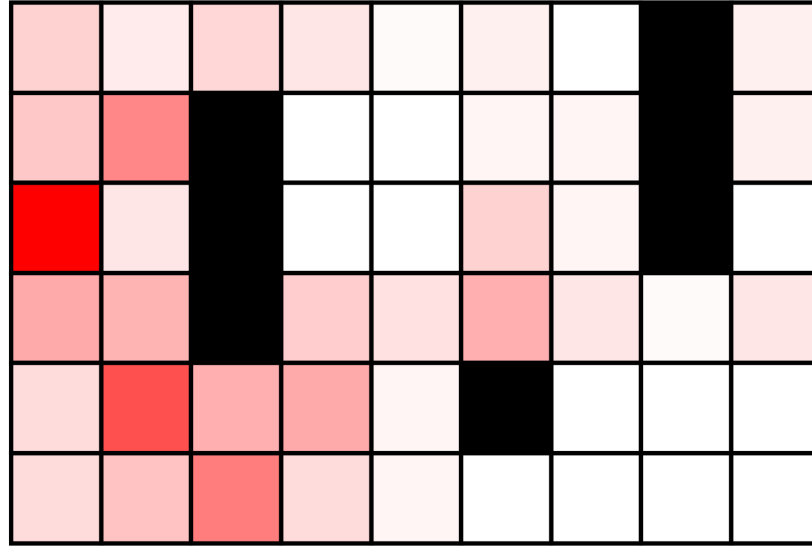


Figure 2.4: Location of the agent when it decides to replay during a sample run of the original Gridworld environment. Darker red indicates more replay activity. Given that replay is more advantageous in certain states than others, we correspondingly predict that agents will feel more fatigued in some states than others.

A Normative Lens to Understand Psychiatric Illnesses

Cognitive control and its attendant costs have been an important focus in the emerging field of computational psychiatry, which aims to give precise mathematical characterizations of mental illnesses in order to increase our understanding of these illnesses and move towards developing effective therapeutics [142, 211, 92]. Cognitive control is often considered to be disrupted in many psychiatric illnesses [40, 29], and much of this work has centered around the notion of control costs. Yet, because our understanding of control costs has been so limited, why they are implicated in psychiatric disorders remain unclear. Framing the costs of control as opportunity costs, and specifically by registering the value of replay as an opportunity cost, may therefore be useful in the effort develop a concrete and normative understanding of psychiatric illnesses. Consider post-traumatic stress disorder (PTSD) as an example. If trauma is associated with an event that elicits a particularly large negative prediction error, rational models of memory sampling [130, 120] and the current need/gain framework suggest that these should be sampled repeatedly and more often than non-traumatic

experiences. This sampling procedure may correspond to the behavioral phenotype of reliving traumatic experiences and rumination. Thus, some symptoms of PTSD may be built on a rational response to negative events, from a rational underlying algorithm being met with an anomalous event [9, 113, 69].

Hippocampal Lesions

One potential challenge to the EVB_C model would be a hypothetical finding that hippocampal-lesioned patients experience and/or exhibit cognitive fatigue. Although there has not been any systematic study of which we are aware that has measured cognitive fatigue in hippocampal-lesioned patients, it seems likely *prima facie* that this would be observed. To the extent that replay is dependent on the hippocampus, the observation of fatigue in the face of damage to this structure would seem to run counter to the model.

Nevertheless, there are two reasons why this might still be observed. One is the possibility that there are multiple offline processing mechanisms, some of which are hippocampal-dependent but some of which are not, in which case fatigue and the benefits of rest might still be observed even in the absence of the hippocampus. This is quite likely. While we developed our theory referencing hippocampal replay in spatial navigation, which is the case with the most relevant experimental detail, there is a longstanding debate about the extent to which these phenomena are specific to navigation vs. a case of a more general function [37]. Moreover, for other tasks in which deliberative planning has been documented (e.g., multiplayer games and rodent instrumental conditioning) there is at best conflicting evidence of hippocampal dependence (e.g., [42]).

The second reason is that, whereas the actual execution of replay may depend (even, for the sake of argument, entirely) on the hippocampus, its engagement is presumably under the control of frontal mechanisms responsible for both monitoring and evaluating the need for replay [96, 192, 132], and inducing it when needed. This would fall squarely within the scope of theories that suggest frontal structures such as the anterior cingulate and dorsolateral frontal cortex are responsible, respectively, for calculating the expected value of control-dependent processes and engaging those deemed to be most valuable [190]. In that case, whereas a lesion to the hippocampus might impair the ability to *carry out* (and thereby benefit from) replay, it may leave intact the ability to *assess the value* of replay, and the phenomenological correlate of the decision that it is worthwhile (i.e., fatigue). This suggests the intriguing possibility that a double dissociation could be found between distinct contributions of hippocampus and frontal cortex to fatigue.

Physical Fatigue

A natural extension of our work is to bring this framework into the domain of physical effort and fatigue. Admittedly, the semantic similarities between cognitive and physical fatigue do not necessarily imply a mech-

anistic similarity (in fact, some have argued that this perceived mechanistic relationship between physical and mental fatigue have been a distraction; [19, 88]). There are clear physiological components to physical fatigue, which are beyond the scope of the current theory. However, there still may be mental and/or motivational components [128]. Whether the effects of these factors are analogous to the role of mental simulation in cognitive tasks may be an exciting direction for future research.

The $EV B_C$ model presented above provides a normative and mechanistic account of the relationship between task difficulty and the dynamics of task engagement, in which fatigue is proposed to signal the value of replay relative to overt task performance. This scales with the difficulty of the task, such that fatigue increases and engagement diminishes with greater difficulty. However, diminishing engagement is not restricted to difficult tasks; it is also observed in easy and/or repetitive ones when they are performed for sufficiently long periods of time. This is commonly associated with another phenomenological experience: boredom. That is, after performing even an easy task for enough time, people often experience boredom and prefer to switch to a new task.

2.5 Arousal and Boredom

As with fatigue, there is a longstanding literature on the phenomenology of boredom, in which it has been argued that people strive for ‘optimal arousal,’ a state in which stimulation is regulated in order to achieve maximum performance [222]. Optimal arousal theory initially focused on arousal associated with purely environmental stimuli, but subsequent work has suggested that optimal stimulation is also dependent on the individual. For example, ‘flow’ has been described as a state in which an individual is voluntarily and fully immersed in their work [44]. However, virtually all of this work has been descriptive; there has been little formally rigorous, theoretical work that explains the purpose of arousal.

One line of work that has pursued both a mechanistic and normative account of arousal has focused on its association with norepinephrine (NE) function – a neuromodulator that is widely distributed throughout the brain. The Adaptive Gain Theory (AGT; [15]) of NE function provides a mechanistically explicit account of the non-monotonic effects of arousal on task function, and a possible foundation for a normative account of boredom. Low levels of arousal (and, in the limit, sleep) are known to be associated with the lowest levels of NE release. According to AGT, moderate states of arousal correspond to conditions under which NE release is increased modestly, but timed to selectively favor processing information associated with the current task improving performance; while the highest states of arousal correspond to higher levels of sustained NE release that are less discriminate relative to task-relevant processes, in effect adding noise to the system and thus degrading performance of the current task (see [15] for the proposed mechanisms of action of

NE and their effects on task performance). AGT argues that the noise induced by these higher levels of NE release, which otherwise might be viewed as a distraction, facilitates switching to other tasks and thus promotes exploration. Furthermore, AGT proposes that this occurs when estimates of task-related reward diminish, and/or the likelihood of other opportunities for reward grows. A closely related model of NE function suggests that it is responsive to increases in uncertainty rather than diminishing relative reward [10]. In either case, or perhaps in both, high levels of NE can be seen as favoring disengagement from the current task set in favor of switching to some other one. There is mounting evidence in support of this idea, both from direct physiological recordings of locus coeruleus (LC; the brainstem nucleus that is the source of NE; [102]) and from the use of pupillometry as an indirect measure of LC function [160].

Here, we consider that boredom is the phenomenological correlate of such signals, and in particular that it indexes the value of exploration (investigating new opportunities that may lead to greater reward in the future) over exploitation (pursuing known, more immediate sources of reward). Mirroring our account of cognitive fatigue, although boredom reflects the *value* of improving future choices, it is perceived as a *cost* disfavoring status quo exploitative action; that is, it indexes the opportunity cost of foregoing the exploration. As discussed below, the value of information from exploration in the real world (as opposed to the value of mental simulation) captures the remaining puzzles of the relationship between task difficulty and task disengagement.

2.5.1 Empirical Findings

To motivate our model of boredom, we first summarize the two primary empirical findings it is meant to explain: 1) boredom is minimized when agents are at an optimal participant-task fit (i.e., the state of “flow”), which is often at an intermediate level of difficulty; and, closely related to the first, 2) if agents are bored, they will seek other tasks to perform.

Optimal Participant-Task Fit

A survey by Fisher [63] found that the causes of boredom in work environments were quantitative underload, qualitative underload, and qualitative overload. In quantitative underload, workers do not have enough to do; in qualitative underload, the tasks are too easy; and in qualitative overload, the tasks are too challenging. Conversely, boredom is minimized during states of ‘flow’ [43, 44, 95, 163, 148], the state of being completely immersed and focused in a task.

A recent set of experiments by Geana et. al [70] directly connected boredom to learning, information gathering, and exploration. In the first of those experiments, participants were presented with a series of

randomly selected numbers from 0 to 100, one at a time, and simply had to predict the next number that would appear. The task was performed in three conditions: in the ‘Gaussian’ condition, numbers were sampled from a Gaussian distribution with a fixed mean and standard deviation; in the ‘Random’ condition, numbers were uniformly sampled between 0 and 100; and in the ‘Certain’ condition, numbers were generated as in the ‘Gaussian’ condition, but participants were told the sampled number before they had to respond, rendering the task trivial. The experiment tested the idea that boredom reflects decreasing information content over time. In that experiment, participants were periodically asked to rate their boredom, and the authors found that this measure was inversely correlated with changes in prediction errors, a proxy for the amount of information being acquired in the task the dynamics of which, in turn, differed between conditions according to what could be learned. A similar relationship between prediction error and boredom has been measured in Antony et. al [11].

Switching to Other Tasks

Rationality dictates that if an agent is not in an optimal participant-task fit, they should seek to switch to other tasks in which there is such a fit. The second experiment of Geana et al. [70] sought to test this, by allowing participants to switch voluntarily among the tasks used in their first experiment. They reasoned that if boredom signals the value of exploration, then it should be possible to demonstrate that participants are willing to forgo reward (i.e., pay) for the opportunity to gain information, by switching to a task that pays less but provides more information. Consistent with this prediction, they found that participants spent the most time in the ‘Gaussian’ condition, in which there was the greatest information content. This behavior runs counter to a standard rational agent model based exclusively on reward, since the ‘Certain’ condition, not the ‘Gaussian’ condition, was the one that maximized reward. This behavior can also be seen in human work environments. Survey results have demonstrated that, to mitigate levels of boredom, many workers perform auxiliary tasks such as reading novels or writing letters [63]. In the extreme, boredom has been found to lead to a higher labor turnover [215, 71, 108].

Finally, [70] conducted a third experiment to test the extent to which opportunity costs associated with task context had an effect on boredom and exploratory behavior. In the first part of the experiment, participants performed a standard two-armed bandit task [24] that was used to evaluate their bias toward exploration, during which they also periodically evaluated their boredom. This was followed by an auxiliary task that was known to the participants up front, and was manipulated across individuals to determine the extent to which knowledge of it had an impact on boredom and exploratory behavior in the bandit task. They found that participants anticipating the more interesting auxiliary task reported the bandit task to be more boring and that

these self-ratings of boredom correlated with increased exploratory behavior in the bandit task. Interestingly, in this experiment, participants could not voluntarily switch from the bandit to the auxiliary task; thus, taking only that particular situation into account, the value of the auxiliary task should not have had any objective effect on the bandit task. What the results suggest, however, is that boredom may reflect the potential value of exploring alternatives even when these are not immediately or obviously accessible (and, conversely, the opportunity cost of not being able to do so). Taken together, the results of these experiments suggest that the experience of boredom accompanies the propensity to explore, and are consistent with the hypothesis that, more specifically, it signals the estimated expected value of doing so.

Discussion

The proposal that boredom signals the value of exploration fits well with the idea that the allocation of control to a given task incurs an opportunity cost with respect to others, coupled with the reasonable assumption that the world is non-stationary – that is, that opportunities for reward change over time. Once one has mastered a task – that is, one has full knowledge about it, and therefore it has become easy (or as much so as possible) – little remains to be learned that might be of use more generally, making it less valuable to continue and more valuable to move on. This idea has been formalized in models of reinforcement learning [182, 159, 219]. In the section that follows, we generalize these formalizations and evaluate the extent to which it contributes to patterns of modulation of performance.

2.5.2 Formalizing the Value of Information

An approximation to the value of information (VOI) can be expressed generically in a form analogous to EVB in the M&D replay model described in Part 1, providing a formal, integrated framework for investigating potential relationships between boredom and fatigue, as follows:

$$VOI(s_k, a_k) = Gain(s_k, a_k) \times Need(s_k, a_k) \quad (2.15)$$

in which

$$Gain(s_k, a_k) = \mathbb{E} \left[\sum_{a \in A} [\pi_{new}(a|s_k) - \pi_{old}(a|s_k)] q(s_k, a) \right] \quad (2.16)$$

$$= \int_{r \sim R(s_k, a_k)} p(r) \left[\sum_{a \in A} [\pi_{new}(a|s_k) - \pi_{old}(a|s_k)] q_{new}(s_k, a) \right] dr \quad (2.17)$$

and

$$Need(s_k, a_k) = \sum_{i=1}^{\infty} \gamma^i \delta_{s_{t+i}, s_k} \quad (2.18)$$

$$= \gamma \sum_{s' \in S} T(s_k, a_k, s') M(s', s_k) \quad (2.19)$$

in which M is the successor representation [51, 75]. The *Gain* term captures the informational value of the obtained reward r , as for *EVB* in terms of the resulting change in the choice policy at state s_k . Here this gain will be realized in expectation over which reward is in fact obtained (i.e. over the prevailing prior distribution of r). Such gain is obtained following every subsequent visit to s_k , as captured by *Need*.⁶

By analogy with the previous section, we identify boredom with the drive to explore, from VOI. We first model task switching via a simple “bandit” task, with one state s (visited repeatedly) and several tasks a , the values $r(s, a)$ of which can change. Then we model boredom as the VOI for the most informative option relative to the current one: $\max_k VOI(s, a_k) - VOI(s, a_t)$.⁷

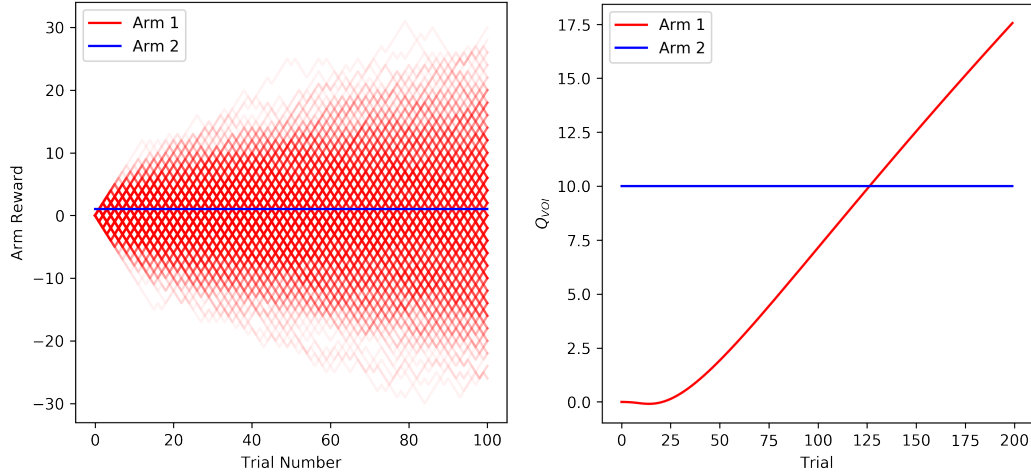


Figure 2.5: Sample bandit task illustrating the increase in Q_{VOI} with increased uncertainty. (Left) Red indicates sample reward trajectories of the first arm, the reward for which starts at zero and then increases or decreases by one every iteration. The blue arm 2 indicates an arm with value always at 1. (Right) The change in Q_{VOI} over time if the agent keeps on picking the blue arm. After a while, the uncertainty of the red arm is large enough that it is advantageous to pick it, even though its expected value is less.

As an illustrative example, consider a simple two-armed bandit. The first arm starts with a value of zero but, with every iteration, it has a fifty percent chance of increasing by one and a fifty percent chance of

⁶These expressions give a partly myopic approximation to VOI: Although they measure the expected future value of exploiting any learning over repeated future visits to s_k , they do not consider the additional informational value of additional learning at these subsequent steps. This approximation was chosen to match the same simplification as used for *EVB* by M&D, as in the previous section, and is sufficient for our purposes.

⁷An alternative characterization may model boredom directly through the Q_{VOI} , i.e. $\max_k Q_{VOI}(s, a_k) - Q_{VOI}(s, a_t)$. We leave the differences between these two formalizations as an important direction for future research.

decreasing by one. The reward is deterministic based on the arm’s current value. The second arm serves as a baseline and always has a value of one. Here, each arm represents a task, and the stochastic dynamics of the first arm embody a simple form of volatility in the value of options in the world.

Let us assume, as a start, that the agent begins by always choosing the second arm, that is, the stationary option. This allows us to see how the VOI for the alternative, dynamic option changes over a prolonged experience with the stationary one. We assume the values for the two actions, Q_{VOI} are tracked (as distributions) using simple, recursive Bayesian inference over the true dynamic model of the task.

Figure 2.5 shows how the action’s Q_{VOI} values change over time. Notice that, even though the expected one-step reward of both arms remains constant (the first at one and the second at zero), the Q_{VOI} of the first arm increases over time. As the trials go on, the uncertainty about the first arm’s value increases — as does the VOI for resolving this uncertainty — and choosing it once can give valuable information about which arm to choose in subsequent trials. If the first arm’s actual reward is higher than the second arm’s, the agent will change policies after exploring. If it is less, the agent simply goes back to its original policy. At some point, when the uncertainty becomes large enough again (i.e., after choosing the second arm for a sufficient number of trials), the first arm becomes a better choice because of the additional VOI, even though its expected reward (i.e., the mean base Q without considering information) is still lower.

The graph in Figure 2.5 captures an observation reported extensively in the boredom literature: as one repetitively does a task, the relative value of alternative tasks increases. In our model, this is due to an increase in uncertainty – and therefore the VOI – about the value of the other tasks. This is also coupled with low uncertainty about the status quo task, for which in the current example VOI is always zero because the task is maximally uninformative and static.

The VOI is nonzero when the Q -value posterior is different than the Q -value prior. The increase in value associated with Arm 1 in Figure 2.5 reflects a situation in which there is learning such that the posterior is different than the prior. In contrast, the value of Arm 2 does not change, since the prior and posterior are never different, and thus there is no gain. This captures the scenario in which a task is too easy, as in Geana et al.’s Certain reward condition. The same thing happens for a task that is too hard, e.g. rewards are stochastic but teach you nothing, as in Geana et al.’s Random condition in which there was no information to be gained.

The foregoing simulations describe dynamics of boredom related to uncertainty with respect to its effects on the Gain term. Our formulation also suggests boredom can be affected by the Need term. This can be examined, for instance, by manipulating the horizon in multi-armed bandit settings. For example, Figure 2.6 plots the probability of choosing Arm 1 in the example above (assuming a softmax decision function), for games of fixed but differing lengths (i.e., numbers of trials), mirroring an effect reported by [218]: participants explore more in games with longer horizons, when they have more opportunities to gain information about

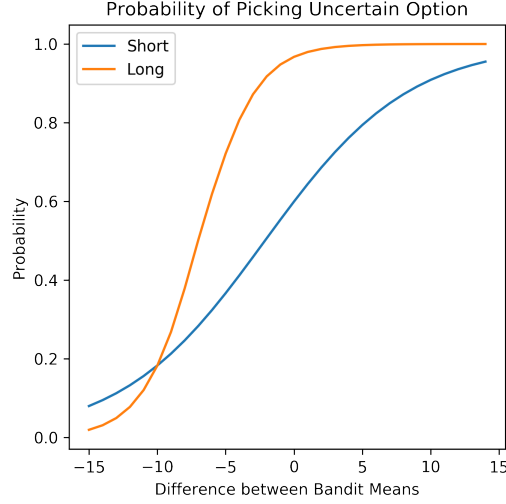


Figure 2.6: For various mean differences between the two bandit arms, we plot the softmax probability of choosing the option with higher uncertainty in the short horizon versus the long horizon. We reproduce the effect found in Wilson et. al [218], in which it was shown that participants in longer horizons are more exploratory. Our formulation provides a parsimonious reason why: the Need term is higher in a longer horizon, and thus the VOI is higher.

the uncertain option. In our formulation, this is because the Need term is higher for longer horizons (more expected future choices in which to exploit any learned policy improvements), and thus the overall value of information is higher.

2.5.3 Discussion

We propose that boredom reflects an adaptive signal used to promote exploration and ultimately achieve higher long-run returns. A large amount of evidence has suggested that boredom arises from a suboptimal fit between the participant and the task. Here, we give a formal argument why a suboptimal fit may not maximize long-term reward: when a task permits learning, its true value should reflect additional future gains due to that learning. The optimal task for an individual, then, is one that is rewarding as well as one in which they can continue to learn. Specifically, the agent should take the action with the highest Q_{VOI} , which balances both the currently known return and the expected increase due to further learning. When these are high, we suggest this corresponds to a *flow* state. This view captures the nonmonotonic relationship between task difficulty and boredom: qualitative underload and overload (tasks that are boring due to being too easy or too hard) both correspond to situations in which the task’s VOI is low. Furthermore, when the agent is bored, i.e. engaged in a task with a suboptimal Q_{VOI} , the agent should try to find a task that is a better fit. Below, we discuss different literatures that our formulation can potentially connect to.

Adaptive Gain Theory

While our analysis has primarily been at Marr’s vision computational level, a complete theory would account for effects at all levels of analysis. As discussed before, [15] offered the ‘adaptive gain theory’ of the locus coeruleus-norepinephrine (LC-NE) system, in which they suggested that phasic LC activation supports exploitation while tonic LC supports exploration. Recently, [102] presented causal evidence supporting this hypothesized relationship. In this work, the authors found that increasing tonic NE levels in LC led rodents to explore more in a patch-foraging task. NE levels provide a mechanism by which the brain implements the value of information computation we proposed. Thus, one fruitful line of work would be to link phasic and tonic NE responses to specific exploration algorithms and determine how they change with boredom, in a similar manner to the connection between dopamine and temporal difference learning [185, 45].

Curiosity

We have focused on boredom as a negative reflection of the value of information, which drives choices away from uninformative options and hence toward exploring more informative ones. It is clearly also the case that there are also affective states associated in a positive way with the informational value of exploration. Although we have focused on flow, another, which is evidently more directed toward individual alternatives, is curiosity. For instance, [58] proposed a model, similar in some ways to ours, interpreting curiosity as an affective state that signals the value of exploring stimuli in terms of their potential for increasing future reward.

The emergence of these formal models will hopefully permit future work aimed at determining the extent to which curiosity vs. boredom are simply mirror images, or instead more or less engaged in different circumstances or reflect different aspects of VOI. For instance, we have emphasized the involvement of boredom in driving switching between tasks, although the same informational considerations (and the same models and algorithms) equally apply for within-task learning as well. Are curiosity and boredom engaged at different hierarchical levels of tasks and subtasks, or equally across them? As another example, boredom most obviously relates to the extent of experience with an option, which is a key determinant of VOI. However, a different signal of VOI — novelty — is often invoked in analyses of curiosity. Do boredom and curiosity actually differ in terms of which features of VOI they are sensitive to? Finally, we have argued (and [70]’s results support) that boredom is a *differential* signal of VOI, in the sense that it is increased not just by (low) VOI for the current task but also by (high) VOI for alternatives. Is this type of symmetry also true for curiosity?

2.6 Mind-Wandering

So far, we have emphasized two different ways in which cognitive or physical actions can be valuable due to their potential for improving one’s future choices. The first part modeled how internal computations — replay — can improve learning by propagating knowledge about rewards to distal states and actions. The second part modeled the physical actions that gather such knowledge in the first place, and the value of visiting informative (e.g., unexplored) states. We treated these mechanisms as separate and parallel, but they can also interact in important ways. One point of interaction, already reflected implicitly in the dynamics of fatigue in Part 1, is that the value of internal replay is ultimately fed by the gathering of actual information (in the external world) to propagate. However, in many tasks, replay could be used to propagate not only the value of known rewards (e.g., as in our previous simulations, to find the best paths to rewards once they are discovered), but also to propagate the value of information. In fact, one reason exploration in more general sequential tasks (like the gridworlds of the first part) is more difficult than in bandit tasks (like the second part) is that in the former, opportunities to obtain VOI can occur at distal states. Reaching those states so as to harvest the VOI itself requires planning, much like figuring out paths to known rewards. Here, we examine the possibility of replay propagating VOI in a combined model, and show that this interaction predicts dynamics of behavior that are different than when replay and exploration are treated separately.

2.6.1 Simulations

Here, we consider simulations of an agent in a T-maze setting. In this environment, there are two terminal states, each of which has its own reward. Furthermore, the environment is non-stationary, such that for every n trials the rewards are shuffled between the terminal states. Thus, the agent needs to continuously explore to adapt to the changing reward structure, and must use replay to plan sequential trajectories both to explore and to exploit these rewarding states.

Model

We augment the EVB_C model from Part 1 by assuming it propagates gain based on $Q_{VOI} = Q + VOI$ rather than reward value Q alone. For the current purpose, we also substitute a different approximation for gain in terms of Q , based on prioritized sweeping [162, 143]:

$$Gain(s_k, a_k) = \gamma^\tau \rho |R(s_k, a_k) + \gamma \max_a Q_{VOI}(s_{k+1}, a) - Q_{VOI}(s_k, a_k)| \quad (2.20)$$

which is the discounted absolute value of the Bellman residual (reward prediction error) at each state. This can be shown to provide an upper bound on the gain as defined previously. This is useful here because, by overestimating gain, it tends to counteract underestimation due to another approximation in our framework.⁸ We also include an optional degree of freedom $\rho \leq 1$ to scale the heuristic, though we set $\rho = 1$ for our simulations. The term involving γ discounts this increased reward by the time it takes to replay. The agent’s opportunity cost by replaying during state s is

$$(1 - \gamma^\tau) \sum_a \pi(a|s) Q(s, a) \quad (2.21)$$

Thus, the agent should replay the state-action pair with the highest $EVBC$ as long as that is greater than the opportunity cost. Once it falls below, it should act.

Finally, we define a new heuristic for VOI appropriate to the dynamics of reward and resulting uncertainty in these environments (i.e., the rewards shuffle every n trials):

$$VOI(s) = U_0 \times (1 - e^{-kN_t(s)}) \quad (2.22)$$

in which U_0 is the maximum value of the uncertainty, k is a constant reflecting the hazard rate for switching, and $N_t(s)$ is the number of trials since the last visit to the state. $N_t(s)$ is initialized as ∞ , meaning the VOI is initialized at U_0 and then drops to 0 once visited. After being visited, it exponentially ‘decays’ back to U_0 , reflecting the accumulating chance that a change will have occurred.

Results

We ran simulations of the agent described above in a T-Maze with different hazard rates (Figure 2.7). The agent develops uncertainty about the rewarding terminal states, and then uses replay to propagate the corresponding value of information back to the initial states. As a result, we see extensive replay at the beginning, when the agent knows about the existence of (but not yet the value of) the terminal states and propagates its uncertainty through its model of the environment. Then, the agent’s replay behavior follows an oscillatory cycle as the uncertainty about the non-visited terminal state increases and decreases. This exploration is important because rewards are shuffled among the terminal states every five trials. As Figure 2.7 demonstrates, an agent with this exploration bonus can use replay to achieve a high average reward rate.

⁸Underestimation arises because $EVBC$ is defined myopically: it fails to account for the value of a single replay operation in permitting subsequent steps of replay. Accordingly, another way to mitigate this problem would be to extend the original replay model to allow n -step backups and calculate their value accordingly.

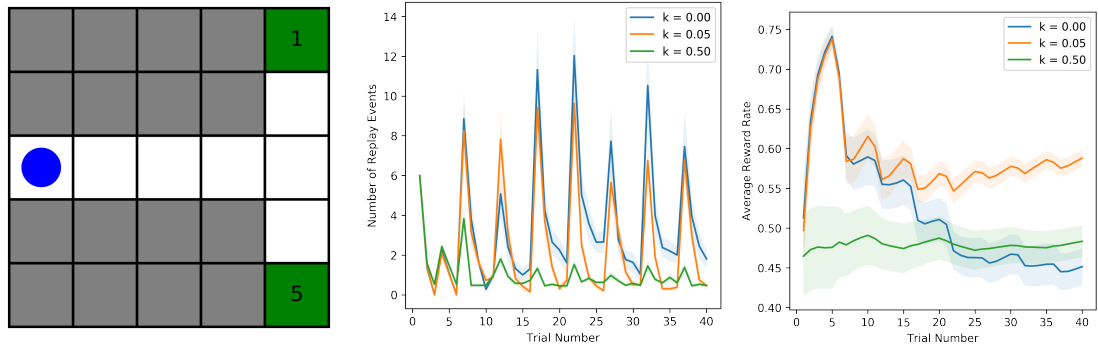


Figure 2.7: Simulations of an agent that replays and explores. The agent uses replay in order to propagate value of information through the agent’s model of the environment. (Left) Gridworld T-maze environment. (Middle) Replay behavior for different hazard rates. (Right) Average reward rate for different hazard rates. Reward rate is greatest for a moderate hazard rate. All error bars indicate ± 1 SEM.

2.6.2 Discussion

It has sometimes been questioned whether performance decrements following a long time-on-task are a result of fatigue or boredom [126, 161]. Meanwhile, other work has not discriminated between these states, implicitly considering both as reflecting motivation and/or opportunity costs [88, 115]. Although the focus of the first two parts of this article was to clarify this distinction, at least hypothetically, by treating them independently of one another, the interaction between replay and exploration considered above shows that there is still ambiguity. In particular, the current simulation predicts two types of replay not present in Part 1: at the beginning of a task, and perpetually even after the task is overtrained. Both of these are due to uncertainty and VOI now driving replay as well as exploratory action. This remains perpetually high because we have assumed (as in the boredom modeling) that the value of terminal states may change if they are long unvisited. Since these relate to both EVB and VOI, it is unclear within our framework which phenomenological state to associate with each.

It is also possible that such interactions could be associated with other phenomenological states. Above, we discussed curiosity as one possibility. Another is ‘mind-wandering,’ the act of spontaneously generating thoughts, that has become a focus of empirical study given its prevalence in daily life [195, 67, 65, 196, 35]. Recent work has suggested that mind wandering can be goal-directed, and can facilitate creativity [224, 216, 66, 2]. While no formally explicit account has yet been offered for mind wandering, our framework suggests the possibility that this might correspond to replaying for exploration — a question that invites future research.

Chapter 3

Behavioral Study

The most common formalization of reinforcement learning, based on Thorndike’s law of effect, directly updates stimulus-response mappings in response to trial-and-error experience. Though optimal, this process can be slow since it requires learning from direct experience. Remote forms of learning, called model-based updating, act as a candidate alternative that can leverage broader knowledge of the environment in order to infer the values of longer-term trajectories.

Hippocampal replay has been recently proposed as one neural substrate underlying model-based computation. Classically, replay involves rodent place cells firing in sequential trajectories, either at choice points or after receipt of reward. Competing accounts argued whether replay reflects planning or memory consolidation, and the M&D model integrated these accounts as value propagation within the framework of reinforcement learning.

In the previous chapter, we extended the original M&D model to be discounted by the amount of reward lost due to the time it takes to replay. This suggestion is in line with costs being attributed to the opportunity cost of time [190, 115], and thus we provided an account of *when* the agent ought to replay. Our $EVBC(s_k, a_k)$ model can be decomposed into the product of *Gain* and *Need* terms:

$$EVBC(s_k, a_k) = \mathbb{E}_{\pi_{\text{new}}} \left[\sum_{i=\tau}^{\infty} \gamma^i R_{t+1} \middle| S_t = s \right] - \mathbb{E}_{\pi_{\text{old}}} \left[\sum_{i=0}^{\infty} \gamma^i R_{t+1} \middle| S_t = s \right] \quad (3.1)$$

$$= \text{Gain}(s_k, a_k) \times \text{Need}(s_k) \quad (3.2)$$

in which

$$\text{Gain}(s_k, a_k) = \gamma^\tau \sum_{a \in A} Q_{\pi_{\text{new}}}(s_k, a) \pi_{\text{new}}(a|s_k) - \sum_{a \in A} Q_{\pi_{\text{new}}}(s_k, a) \pi_{\text{old}}(a|s_k) \quad (3.3)$$

and

$$Need(s_k) = \mathbb{E} \left[\sum_{i=0}^{\infty} \gamma^i \delta_{s_{t+i}, s_k} \mid s_t \right] = M(s_t, s_k) \quad (3.4)$$

We defined $EV B_C^*$ as $\max EV B_C$ and proposed that a rational agent should replay the most valuable location, $\arg \max EV B_C$, as long as its value $EV B_C^* > 0$.

A significant amount of work has demonstrated that the *content* of hippocampal replay is goal-directed [53, 99, 165, 113]. Rewards and paths leading to rewards are some of the most replayed events, and these observations were a central motivation for the original M&D model (the *Gain* for the reward locations and the events leading up to it is high, and this increases the value of the *EV B*). Furthermore, the *Need* term in the M&D model was able to capture why replay is in the forward direction at starting states and in the backward direction at ending states.

Our $EV B_C$ modification additionally proposes that the *use* of hippocampal replay is control-dependent. That is, agents ought to replay more (and subsequently, ‘rest’ more) when replay is more valuable. This mechanistic claim we propose is novel to the hippocampal literature, which implicitly assumes agents replay because they are at rest. Here, we are proposing a novel causal direction, suggesting instead that agents ‘rest’ *because* replay is more valuable than action rather than vice versa.

Our proposal aligns with a growing line of work in model-based reinforcement learning examining the *depth* of computation [106, 189]. While initial work in model-based versus model-free reinforcement learning implicitly viewed these as two extremes on a joint spectrum, recent work has aimed to identify a middle ground in which agents arbitrate to the extent to which they perform model-free and model-based reasoning. The amount of model-based reasoning is adapted to the situation, and manipulations such as increasing stakes and increased cognitive resources tend to increase the depth of computation.

3.1 Experimental Methods

Motivated by these considerations, we designed an experiment that aims to determine the extent to which agents engage in ‘replay’ (or some type of remote / model-based updating) and how this is modulated by terms in the $EV B_C$ equation. Critically, we will invoke a blunt manipulation of the *Gain* (and thus the value of replay; *i.e.* it more or less valuable after certain trials), and we hypothesize the amount of computation will correspondingly change in accord with this manipulation. That is, when it is more advantageous to replay, we expect agents to perform more model-based computation and that this will be validated through behavioral and neural measures.

3.1.1 Background: Two-Step Task

The Daw two-step task [48, 110] is a ubiquitously used experimental paradigm that was developed and refined to isolate model-based and model-free reinforcement learning processes. In this paradigm, agents are initially faced with a binary choice in the starting state, *i.e.* stimulus1 or stimulus2. There are two end states, with stimulus1 leading to end1 70% of the time (and to end2 30% of the time), and stimulus2 leading to end1 30% of the time (and to end2 70% of the time). Each end state gives a binary reward that is parameterized as a bandit with a slowly drifting probability of reward.

The experiment is designed as such to test the behavioral effects of the improbable transition. If the participant selects stimulus2 and experiences the improbable transition to end1 and is then rewarded, model-free and model-based learning predict opposite responses. Model-free learning updates direct stimulus-response mappings, and thus a model-free agent will be likely to select stimulus2 again. On the other hand, model-based reasoning computes the decision tree before action and would realize that in order to return to the rewarded end1, the agent should select stimulus1 as it gives a more probable transition to the rewarding end1. Analyzing the proportion to which participants stay after experiencing the improbable transition thus serves as a proxy for the extent to which the agent recapitulates model-based versus model-free learning.

3.1.2 Task

The study here adapts the MEG reinforcement learning-based experiment in [122], which itself extends the original two-step task [48], to an fMRI setting. The participant starts off in one of three starting states (face, place, object) with (fixed, but randomized per participant) probabilities of 3/6, 2/6, 1/6. (These different starting probabilities correspond to the *Need* term in the M&D model, which will also affect the replay value of each state). In each starting state, the participant is presented with a choice of two images, one that always leads to the final state X and the other that always leads to the final state Y . States X and Y further have a (slightly drifting) probability of reward. Figure 3.1 outlines this design.

The logic of this experiment centers around the fact that three images (one in each category) lead to the final state X and three that lead to the final state Y . Consider the situation in which a participant chooses Face1, leading to the final state X , and then receives a reward. Under the standard model-free reinforcement learning temporal difference update rule, the participant will increase their Q -value for Face1 upon reward. Importantly, if the participant is able to remotely update (model-based reasoning), and if Place2 and Object1 also lead to state X , the agent ought to increase those Q -values in the experiment's allotted time to replay. This experiment is thus designed to gauge to what extent agents are engaged in this type of remote updating.

As discussed before, the main mechanistic claim of our $EV B_C$ model claims that participants would

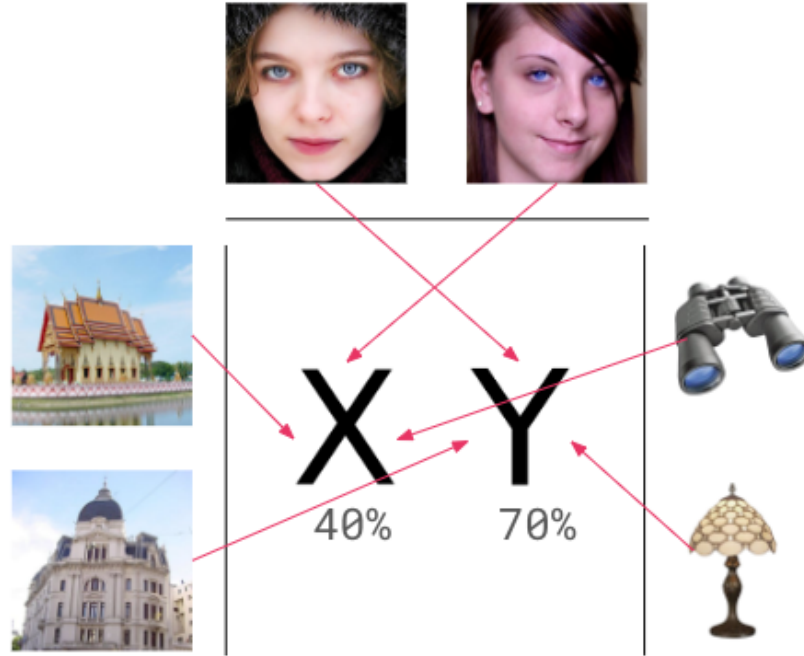


Figure 3.1: In this experiment, the agent starts in one of three starting states (face, place, object). One item in each category leads to end state X and the other to end state Y . End states X and Y have differing (and changing) probabilities of reward, and the participant’s aim is to select the item in each category which will maximize reward.

remotely update more when it is more beneficial to do so (*i.e.* when the *Gain* is higher). Thus, we add an overt manipulation of having 25% of trials being bonus trials (participants are rewarded 5 points instead of just 1). We hypothesize participants perform more model-based computation in the rest periods preceding bonus trials compared to ‘normal’ trials.

Participants are given 200 trials, with the opportunity to take a short break halfway through the experiment. Each trial starts off with a 500ms fixation cross before presenting the first stage stimulus (two faces, two places, or two objects). The participant has 750ms to select a choice (enforcing habitual / model-free decision-making), and then a 1000ms transition period to the second stage. The participant then has 750ms to select the second stage choice (there is only one option so it is not a regular ‘choice’; we include this in the experiment to enforce the participant’s attention), followed by a 1000ms transition period to the receiving the reward (or not). Last, we have a jittered inter-trial interval (1500, 2500, 3500, 4500, 5500, 6500; ms) where we hypothesize remote updating will occur. The experiment timeline is outlined in Figure 3.2.

The type of trial (regular versus bonus) is indicated by a thin rectangular box outlining the experiment. Regular trials are indicated with black lining whereas bonus trials are indicated with rainbow-colored lining. When the participant receives the reward, the next trial type is indicated at the bottom and then that box

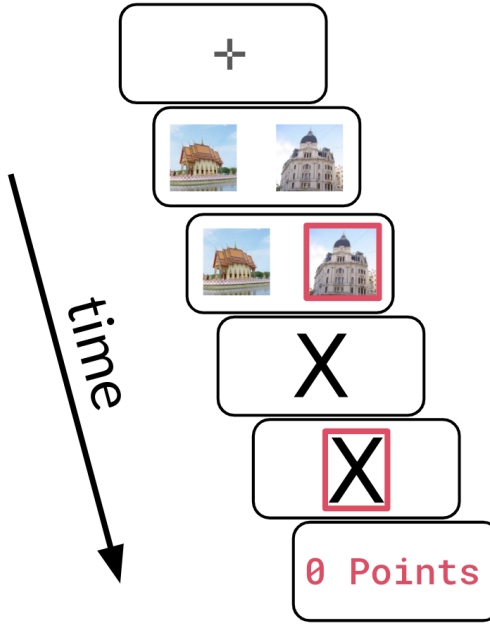


Figure 3.2: Experimental Timeline. Participants first attend to a fixation cross and then select one of two stimuli from a specific category. They are then taken to an ending state, in which they stochastically receive a reward. This experiment is a generalization of the traditional Daw two-step task [48] and an fMRI analog of a recent MEG study [122].

eventually scrolls to replace the current rectangular box, ensuring the participant knows what trial type is pending during the inter-trial interval (when the hypothesized model-based updating is occurring). Thus, we hypothesize increased model-based updating to happen prior to the bonus trial, when the participant has received a reward (or not) from the current trial and has been signaled that the upcoming trial will be bonused.

Before performing the task, participants underwent a ten-minute instruction period, in which they were trained on the structure of the task and performed fifty practice trials. Images and probabilities were different than the ones experienced in the real task, but the underlying structure of the test task remained identical.

3.1.3 Participants

48 participants were recruited from the Princeton community. The sample size was chosen in order to achieve roughly 30 participants after exclusion criteria (excessive motion, satisfactory task performance), which is slightly above the standard in the field today. All participants signed an informed consent form, and this study was approved by Princeton University’s IRB committee.

Table 3.1: Parameters for Logistic Regression Model Fits

Parameter	Mean	95% CI
Intercept	0.30	[0.19, 0.41]
Last Reward	0.30	[0.22, 0.39]
Different State	-0.11	[-0.17, -0.04]
Last Reward x Different State	-0.11	[-0.18, -0.04]

3.1.4 Functional Localizer

Participants completed two equivalent “block-based” functional localizer runs designed to isolate category-specific representations in the brain [173]. Each run contained twelve different blocks, four repetitions of each of the three categories (face, place, object). Each block ran for 40 seconds, with 28 seconds of images (14 images, 1.4 seconds on, 0.6 seconds off), and 12 seconds of rest. To ensure participants were attending to the task, participants were instructed to press a button when an image turned upside down (one per block).

The attention task may have introduced confounds, which will be discussed when we show the neuroimaging results. Some images, specifically faces and places, were easier to identify as being upside down, whereas other images, specifically, objects, were difficult. Activation for the object stimuli may thus not solely reflect object-specific regions, but also regions recruited for the attention task.

3.2 Win-Stay, Lose-Shift (WSLS)

The win-stay, lose-shift (WSLS) strategy is a useful baseline to evaluate participants’ behavior on our behavioral task. Under this strategy, participants repeat their choice (*i.e.* ‘stay’) in the starting state when it leads to a rewarding outcome and change their choice (*i.e.* ‘shift’) when not rewarded. Figure 3.3 analyzes the extent to which participants play this strategy by plotting the probability of a participant staying as a function of two variables: whether the last trial was rewarded or not, and whether the current trial’s starting state is the same or different from the previous trial’s starting state. Model-based learning corresponds to whether an agent can propagate reward information to different starting states and can thus be approximated by the difference in stay probabilities in different states (*i.e.* Diff-Rew - Diff-Non). In contrast, model-free learning corresponds to how an agent updates in the same state and can be approximated by the difference in stay probabilities in some states (*i.e.* Same-Rew - Same-Non) *minus* the model-based effect. Across all participants, agents statistically stay more when rewarded in both the same state ($p < .001$) and a different state ($p < .001$), suggesting that both direct and remote learning are occurring.

Table 3.1 outlines a Bayesian hierarchical mixed effects logistic regression model evaluating the stay probability as a function of reward and state transition. Each participant was modeled as a random effect

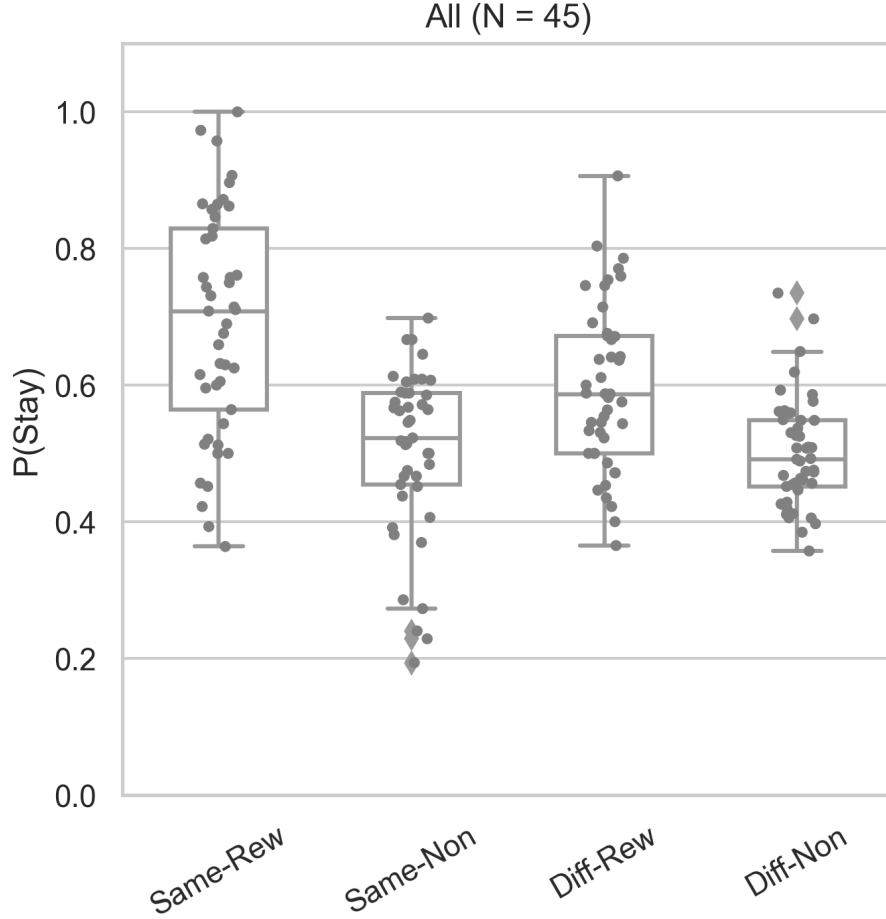


Figure 3.3: Win-Stay, Lose-Shift results for all participants. Same/Diff corresponds to whether the current trial’s starting state is the same/different compared to the previous trial’s starting state. Rew/Non corresponds to whether the previous trial was rewarded or not. The Y-axis reports the proportion of stay, *i.e.* to what extent the proportion repeated their choice from the previous trial to the current trial.

for each parameter, but group means are reported in the Table. Analyses were done in R with the *brms* [30] package, which relies on a Stan [31] backend. Default non-informative priors were used in the analyses.

The model-based effect is calculated by the stay probability when rewarded and in a different state, and the model-free effect is the stay probability when rewarded in the same state *minus* the model-based effect. For Table 3.1, the model-based effect is $0.32 - 0.13 - 0.12 = 0.07$ and the model-free effect is $0.32 - 0.07 - 0.25$. In line with Figure 3.3, these logistic regression results also suggest both direct and remote learning are occurring, with the magnitude of model-free / direct learning being greater than the magnitude of model-based / remote learning.

For subsequent analyses, we consider the baseline level of satisfactory performance to be the agent choosing to stay more when rewarded than not, and we exclude all participants who do not meet this criterion across both state conditions. Figure 3.4 plots the stay probabilities under this exclusion criterion and Table 3.2 re-

Table 3.2: Parameters for Logistic Regression Model Fits (WSLS Subjects)

Parameter	Mean	95% CI
Intercept	0.48	[0.32, 0.64]
Last Reward	0.49	[0.40, 0.59]
Different State	-0.14	[-0.23, -0.06]
Last Reward x Different State	-0.16	[-0.25, -0.08]

ports the results of the Bayesian hierarchical mixed effects logistic regression model. By design, Figure 3.4 and Table 3.2 amplify the learning effect in both the same states and different states, and thus the remote and learning effects remain statistically significant. As a sanity check, the magnitude of model-free / direct learning remains greater than the magnitude of model-based / remote learning.

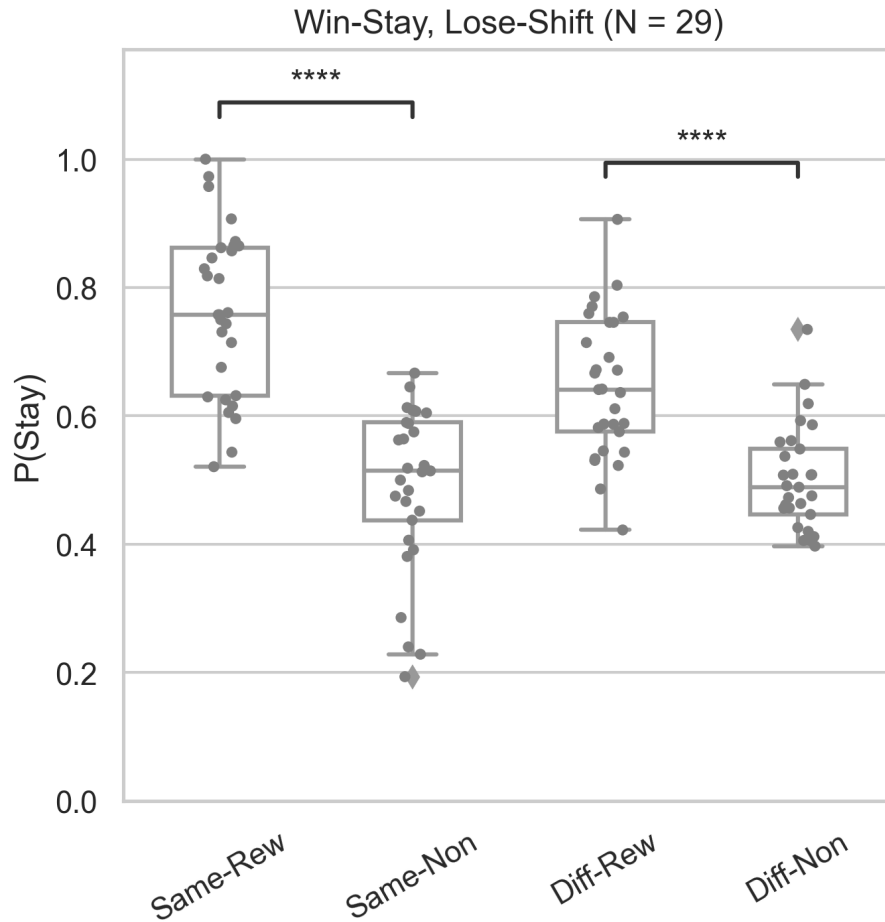


Figure 3.4: Win-Stay, Lose-Shift for participants who stay more across both the same and different conditions.

Lastly, we add a third condition in Figure 3.5, corresponding to whether the current trial is a bonus trial or not. Here, we find that the four first-order effects of direct and remote learning across regular trials and bonus trials are numerically and statistically significant ($p < .001$). We hypothesized there would be a second-order

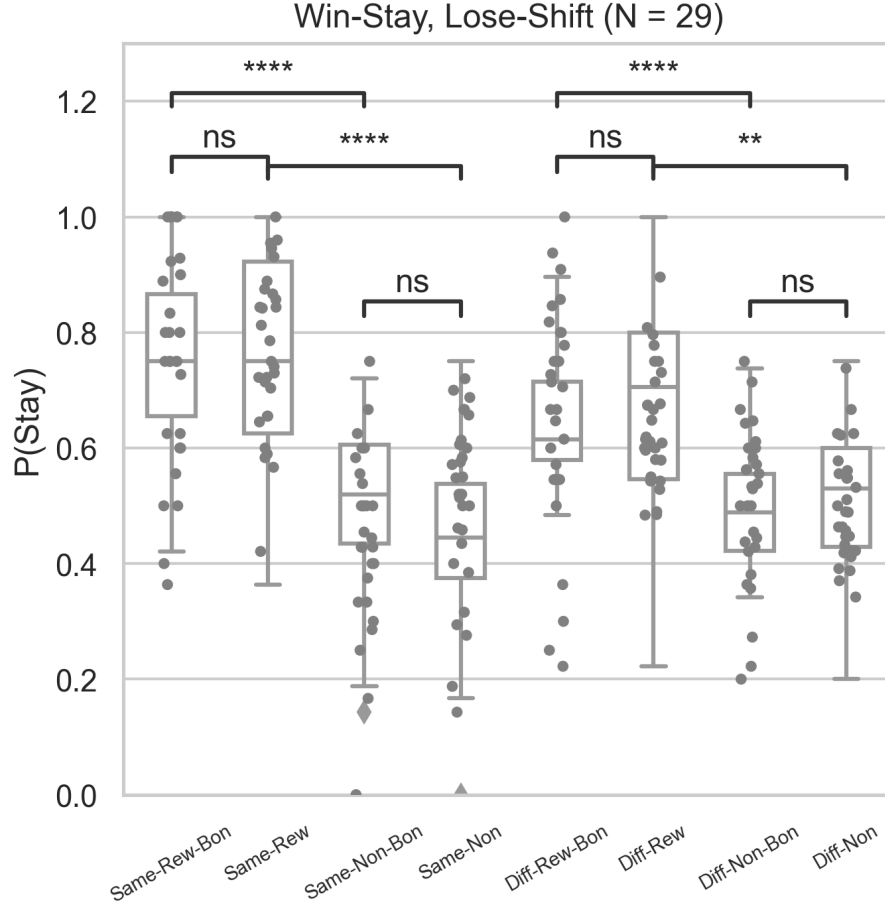


Figure 3.5: Win-Stay, Lose-Shift results for participants who stay more across same and different state conditions, with bonus versus regular trials separated out.

effect of bonus trials, where direct and remote learning would increase in bonus trials compared to regular trials. We actually found learning numerically *decreased* in both local ($-.058$) direct and remote ($-.016$) learning, though none of these results were significant at the $\alpha = 0.05$ level. Despite the fact that these second-order results were in the opposite direction as hypothesized, we do see a potential third-order result in which the bonus trials aid updating more in remote learning compared to direct learning, though this result is not significant at the $\alpha = 0.05$ level. Table 3.3 reports the corresponding Bayesian hierarchical mixed effects logistic regression model coefficients.

Table 3.4 reports learning when the previous trial, as opposed to the current trial, is a bonus trial. This was done to isolate any increased learning effects from receiving a reward of 5 versus 1 (for all previous analyses, the reward is coded as binary even if the participant receives 5 points). Like Table 3.3, none of the bonus parameters in Table 3.4 suggest there is an aggregate learning benefit from bonus trials.

Table 3.3: Parameters for Logistic Regression Bonus Model Fits

Parameter	Mean	95% CI
Intercept	0.48	[0.32, 0.64]
Last Reward	0.52	[0.43, 0.62]
Different State	-0.12	[-0.20, -0.04]
Current Bonus	0.00	[-0.08, 0.08]
Last Reward x Different State	-0.17	[-0.26, -0.09]
Last Reward x Current Bonus	0.05	[-0.03, 0.13]
Different State x Current Bonus	0.03	[-0.05, 0.12]
Last Reward x Different State x Current Bonus	-0.02	[-0.10, 0.06]

Table 3.4: Parameters for Logistic Regression Last Bonus Model Fits

Parameter	Mean	95% CI
Intercept	0.47	[0.32, 0.63]
Last Reward	0.48	[0.38, 0.58]
Different State	-0.12	[-0.20, -0.03]
Last Bonus	-0.01	[-0.08, 0.07]
Last Reward x Different State	-0.15	[-0.24, -0.06]
Last Reward x Last Bonus	-0.02	[-0.10, 0.05]
Different State x Last Bonus	0.04	[-0.05, 0.12]
Last Reward x Different State x Last Bonus	0.03	[-0.05, 0.10]

3.3 Hierarchical Bayesian Analyses

The previous section analyzed behavioral results using the relatively model-agnostic strategy of Win-Stay, Lose-Shift. A more direct way to evaluate the results is to fit hierarchical Bayesian reinforcement learning models. Here, we assume that each agent estimates a value function $Q(s, a)$ for each of the two actions a over the three starting arms s , and the agent chooses with a softmax, i.e.

$$P(a_c = a | s_t = s) \propto \beta Q(s_t, a_t)$$

in which β is the inverse temperature (and a free parameter).

The standard Q -learning update rule in response to reward r after choice a_c is defined as

$$Q(s, a_c) = Q(s, a_c) + \alpha(r - Q(s, a_c))$$

in which α is the learning rate. We create hierarchical Bayesian reinforcement learning models to infer the values of α and β . We model two scenarios. The first is the standard update rule, which we will call direct learning, and denote as when the agent starts in and updates s_d with learning rate α_d . The second is updating the corresponding Q -values for the remote states s_{r1} and s_{r2} with learning rate α_r . In other words, if the participant chooses Face1 which leads to end state Y, direct learning updates the Q -values for Face1, whereas

Table 3.5: Parameters for Next Bonus Model Fits

Parameter	Mean	95% CI
α_d^*	0.37	[0.15, 0.59]
α_d	0.44	[0.28, 0.61]
α_r^*	0.26	[0.10, 0.43]
α_r	0.27	[0.13, 0.40]

remote learning updates the values for Place2 and Object1 (assuming these are the states that lead to end state Y).

Thus, we have three update rules after every trial:

$$\begin{aligned}
Q(s_d, a_c) &= Q(s_d, a_c) + \alpha_d(r - Q(s_d, a_c)) \\
Q(s_{r1}, a_c) &= Q(s_{r1}, a_c) + \alpha_r(r - Q(s_{r1}, a_c)) \\
Q(s_{r2}, a_c) &= Q(s_{r2}, a_c) + \alpha_r(r - Q(s_{r2}, a_c))
\end{aligned} \tag{3.5}$$

Parameters α and β were inferred using Stan [31]. Group means and standard deviations were initialized at $\mathcal{N}(0, 1)$ (with standard deviations having an added lower bound of .001), and each subject mean and standard deviation was sampled from a normal distribution parameterized by the group means and standard deviations. Furthermore, the α learning rates were estimated with standard a normal distribution support of $[-\infty, \infty]$, but then transformed to a support of $[0, 1]$ via the ϕ cumulative normal distribution function.

Our model inferred α_d to have a mean of 0.33 with 95% credible interval [0.17, 0.50] and α_r to have a mean of 0.15 with 95% credible interval [0.02, 0.29]. To determine whether this difference is significant, *i.e.* whether $\alpha_d > \alpha_r$, we directly estimate $\alpha_d - \alpha_r$, inferring a mean of 0.18 with confidence interval $[-0.02, 0.39]$. Thus, we have evidence the direct learning rate is numerically, but not statistically, more than the remote learning rate.

We next determine whether the learning rates are affected by the bonus manipulation, *i.e.* whether the agent replays more when it is more beneficial to do so. In Equation 3.6, we extend Equation 3.5 with an additional three equations that have a bonus-specific learning rate, which we denote as α_d^* and α_r^* . We estimate α^* in two ways.

$$\begin{aligned}
Q(s_d, a_c) &= Q(s_d, a_c) + \alpha_d^*(r - Q(s_d, a_c)) \\
Q(s_{r1}, a_c) &= Q(s_{r1}, a_c) + \alpha_r(r - Q(s_{r1}, a_c)) \\
Q(s_{r2}, a_c) &= Q(s_{r2}, a_c) + \alpha_r(r - Q(s_{r2}, a_c))
\end{aligned} \tag{3.6}$$

The first way is to initialize α^* in the same way as α , *i.e.* with a normal distribution centered at 0 and

Table 3.6: Parameters for Current Bonus Model Fits

Parameter	Mean	95% CI
α_d^*	0.41	[0.25, 0.57]
α_d	0.48	[0.33, 0.62]
α_r^*	0.23	[0.07, 0.36]
α_r	0.33	[0.19, 0.46]

Table 3.7: Parameters for Additive Bonus Model Fits

Parameter	Mean	95% CI
α_d	0.37	[0.23, 0.52]
k_d	-0.16	[-0.78, 0.49]
α_r	0.19	[0.08, 0.32]
k_r	0.09	[-0.40, 0.54]

standard deviation of 1. Table 3.5 reports the results of parameters α_d^* , α_d , α_r^* , and α_r . Surprisingly, we get the opposite effects as predicted. Direct learning decreases with the next trial being bonused, whereas there is no meaningful difference between remote learning.

While Table 3.5 reports the results when the upcoming trial is bonused, directly evaluating our hypothesis that increased stakes will incentivize more model-based computation. For completion purposes, we also model the situation in which the current trial is bonused (or not). Under our formalization, reward is binary (*i.e.* either 0 or 1), even when the actual bonused reward is 5. Thus, it may be more ecologically accurate to model those trials with a higher learning rate (again, assuming that reward is formalized as binary). Table 3.6 reports the results of this analysis.

In both the direct and remote learning cases, learning rates are numerically (but not statistically) greater in the bonus condition compared to the regular condition.

The second way is to initialize α^* as a function of α . We let $\alpha^* = \alpha + k$, in which k is initialized as a normal variable with mean 0 and standard deviation 1. A significant bonus effect would result in a statistically positive value of k . Table 3.7 reports the results of these simulations, in which α is reported after the ϕ approximation, whereas k is denoted with regular $[-\infty, \infty]$ support. We see that k_d is numerically, but not statistically, negative whereas k_r is numerically, but not statistically, positive. Confidence intervals for k are large, suggesting that these results are likely inconclusive.

3.4 Discussion

In summary, we have inconclusive evidence as to whether the use of replay is control-dependent. While some of the hierarchical Bayesian analyses suggest there may be an effect, the effect seems largely small and thus we may need increased power to be confident about the size and significance of the effect.

There are multiple directions going forward for this experiment. First, as hinted, a larger sample size can be collected online, which would hopefully give conclusive evidence to the extent of whether there is an effect or not. Anecdotally, online participants are highly sensitive to incentives and reward manipulations and thus an online study with a few hundred participants may be more informative.

Second, given the seemingly paradoxical effects of the Gain manipulation, it may be the case that the instructions were not sufficiently clear. Participants in [121] pre-trained on the experiment for a day with instructor guidance, and perhaps something similar should have been done for the present experiment.

Last, the proposed experiment doesn't directly test the mechanistic claim as participants do not have the ability to modulate their rest. While we did this for neuroimaging reasons (see next Chapter), a pure behavioral study could let participants directly determine their inter-trial interval and this could be mapped onto 'rest duration'.

Chapter 4

Neuroimaging Analyses

Functional magnetic resonance imaging (fMRI) is a common tool used to noninvasively measure human brain activity. Initial work primarily used it as a tool to identify areas of functional connectivity, in which localized regions are activated in response to specific stimulus categories. For example, a region which is now called the fusiform face area selectively encodes faces [103] whereas the parahippocampal place area selectively encodes scenes [61]. Work in reinforcement learning has integrated fMRI in order to track hypothesized latent constructs at a trial-by-trial level [188, 221, 48, 55]. Model-free reward prediction errors (RPE) are correlated with striatal BOLD activity and the extent to which a participant exhibits model-based versus model-free behavior is correlated with prefrontal BOLD activity.

Replay's fast temporal dynamics have made it a difficult target to study using fMRI methods, which are notorious for their slow temporal resolution (though see [184]). Human replay has been decoded in MEG [122], and our experiment aims to reproduce these effects in fMRI, overcoming the temporal resolution barrier by using a 1 second TR. This aggressive TR may introduce noise and contamination into the BOLD signal, and thus we leverage multi-echo (ME) imaging to reconcile these potential confounds. In contrast to single-echo fMRI which acquires images after a single fixed delay ('echo time'), ME collects three images [168, 166, 114, 125] and combines them posthoc to improve signal quality.

Multi-echo T2*-weighted echo-planar images (EPI) with blood oxygen level-dependent (BOLD) contrast were collected on a 3-T Siemens Prisma MRI scanner. Slices were collected using a multi-band sequence with interleaved slice acquisition with an acceleration factor of 4. Seventy axial slices (3x3x3 mm) were acquired in the oblique orientation of 30 degrees to the anterior-commissure posterior-commissure line with a repetition time (TR) of 1s and 60-degree flip angle, 192 mm field of view. Before the main task, a high-resolution T1-weighted anatomical image (magnetization-prepared rapid-acquisition gradient echo sequence

1x1x1 mm voxels) was also collected. A field map with two EPIs with opposite phase-encoding polarity was also collected, but this has not been integrated with the functionals and anatomicals. Preprocessing was done through fmriprep [62], which leverages the TE-dependent analysis Toolbox (tedana) [59] to perform ME-ICA.

4.1 Localizer Contrasts

Before the task, participants underwent two block localizer runs, with each run containing four blocks for each of the three categories (face, place, object). We used SPM [14] to perform contrasts, combining both runs per subject in a first-level analysis, and then averaging across subjects for the second-level group analysis.

The design matrix contained all the stimuli presented, with a boxcar function of 1.4 seconds per stimulus. This boxcar was then convolved with the hemodynamic response function (HRF) with a high-pass filter of 128 Hz. Motion correction was accomplished by regressing out the six head motion parameters (X, Y, Z, pitch, yaw, roll), which were automatically computed through the fmriprep preprocessing pipeline.

Next, we calculate localizer contrasts for the ‘good’ subjects (*i.e.* those that exhibit win-stay, lose-shift behavior). The three images in Figure 4.1 correspond to one-hot encoded (*e.g.* [1 0 0]) group-level contrasts of face, place, and object regions, activating any individual voxels that meet an uncorrected p-value threshold of 0.01. These contrasts correspond to the activation of these regions versus rest. Cursors are placed in the respective regions of interest (fusiform face area, parahippocampal place area, and lateral occipital cortex). Unsurprisingly, visual cortex activation is high in these specificity contrasts.

To further determine the specificity of these contrasts (and exclude activation of broad regions such as visual cortex), we plot category versus other categories in Figure 4.4. (For example, face vs. other categories is coded as [1 -0.5 -0.5]). In the face versus other categories, FFA activation is surprisingly minimal, though still selective. There is some midline activity, potentially corresponding to a default mode activation, but this activity generally remains small. Recent work [1] has demonstrated familiar facial activation in posterior cingulate, and a related mechanism may explain activation in our paradigm. In the place versus all contrast, activation is strong and robust. Our contrast corresponds well to work in the literature isolating place to parahippocampal place area [61, 18]. Lastly, the object versus all contrast displays robust activation at the edge of the lateral occipital cortex, consistent with historical findings [79, 175]. Unlike the previous place contrast, our object contrast is not well-localized; there seems to be consistent activation in disparate areas across the cognitive control network [41]. One plausible explanation for this finding is our attention task: we asked participants to press a button when an image was displayed upside down (one per block). The contrasts suggest participants may have had difficulty knowing when the object was turned upside down, thus activating

parts of the cognitive control network. Future analyses likely need to isolate the LOC using template atlases.

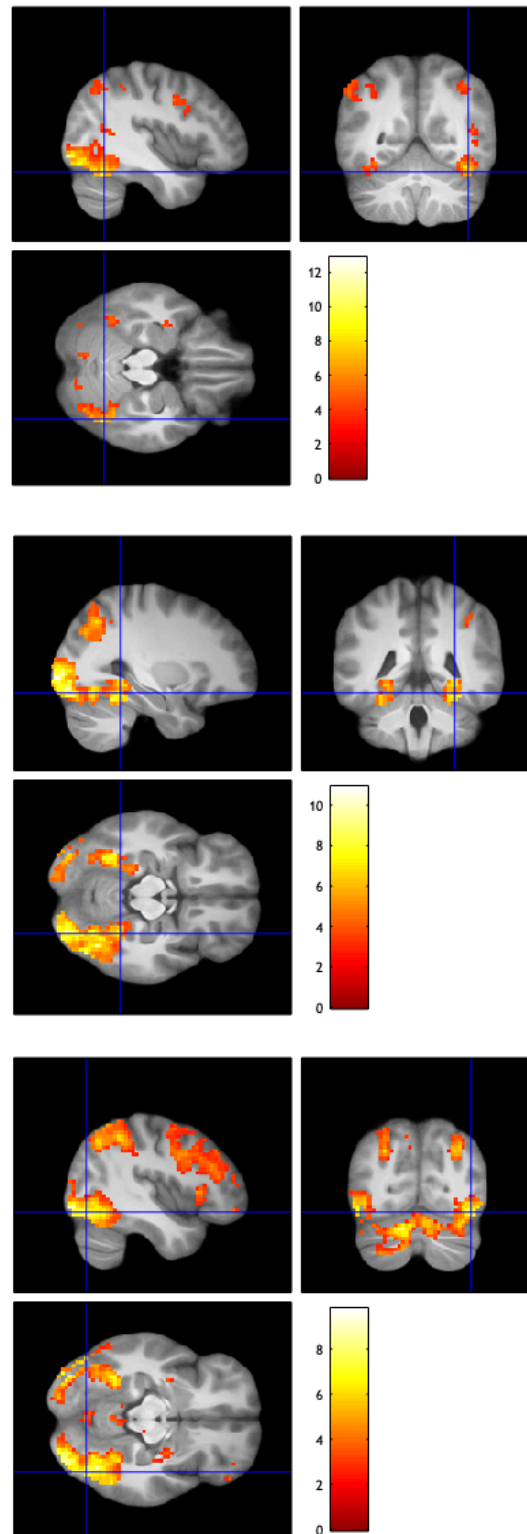


Figure 4.1: Face, place, and object category contrasts. Effects are with $p < 0.01$, and images are plotted in MNI space.

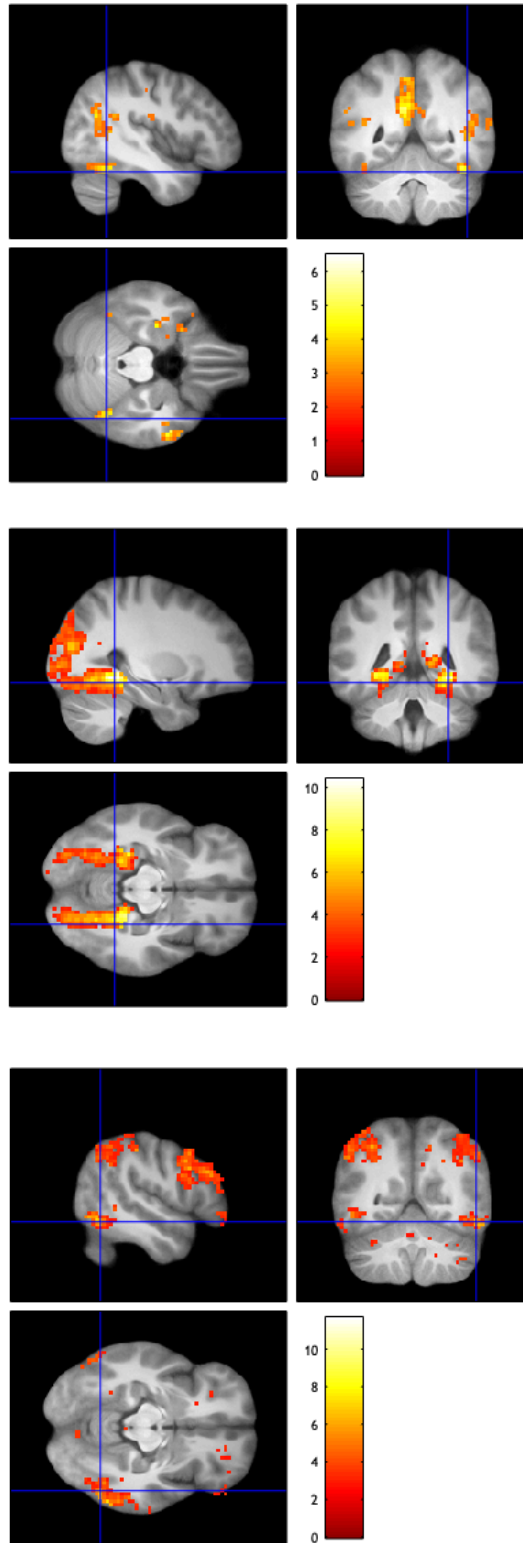


Figure 4.2: Category versus all contrasts for the categories of face, place, and object. Effects are with $p < 0.01$, and images are plotted in MNI space.

4.2 Category Classifier

Now, we seek to create a machine learning category classifier that can decode whether the participant is activating regions associated with faces, places, or objects. We will demonstrate our performance on an individual subject, with the method eventually generalizing to all subjects (aggregate results are reported in the Appendix).

First, as a visual inspection, Figure 4.3 reports the one-hot category contrasts for our sample subject, and Figure 4.4 shows the selectivity versus other categories. Images are in MNI space and thresholded at a cluster minimum of 10 voxels at an uncorrected p-value of 0.001. Initial visual inspection reveals face activation in the fusiform face area, and object activation in the lateral occipital cortex, but no immediately visible place activation in parahippocampal place area. We will discuss these results when analyzing classifier performance.

To process the data for our classifiers, we z-score the BOLD activation in a subject's native space, and restrict the search space to specific ROI's defined from the Harvard-Oxford atlas (which we also transform from MNI to subject native space). We use a support vector classifier with a linear kernel to predict category activation from the z-scored data. A 3-5 TR offset is used to map between stimulus and response data, with the specific offset being chosen to maximize out-of-sample predictive accuracy.

Table 4.1 reports the binary classification of each of the three categories versus rest for different ROI subsets (with Table 4.2 listing all the ROIs). Peak performance is roughly 90% (chance is at 50%), with performance being maximized at the most inclusive subset of ROIs. The results in this table collapsed data across both runs and used a 5-fold cross-validated procedure. While it is difficult to understand which ROI(s) are the most useful for our classifiers, activity in LOC (ROIs 23 and 24), as well as parahippocampus (ROIs 35 and 36), seem to be highly selective in separating rest from category activation, consistent with their historically measured roles in face, place, and object activation.

Table 4.3 similarly reports the binary classification of each of the three categories versus rest but uses a different training and testing procedure. Here, we train a classifier on one of the localizer runs and test it on the other run, thus evaluating the ability to transfer between runs. We do this twice for both combinations of the two runs, and report the average of the train and test accuracies. Again, performance is roughly 90% (chance is at 50%), suggesting there is an adequate signal for more complex classification models.

Tables 4.4 and 4.5 report four-way classification performance within and across runs. Chance is at 25%, but performance peaks at roughly 66% in the across-run condition with the most inclusive ROI's. While ideal performance would be upwards of 90%, a close-to-70% performance is likely sufficient to start decoding replay events during rest periods.

Table 4.1: Binary Classification of Category vs. Rest (Within Runs)

ROIs	Category	Train Accuracy	Test Accuracy
22	Face	0.838	0.757
22	Place	0.871	0.777
22	Object	0.854	0.767
23, 24	Face	0.927	0.894
23, 24	Place	0.935	0.901
23, 24	Object	0.901	0.856
35, 36	Face	0.922	0.882
35, 36	Place	0.927	0.890
35, 36	Object	0.908	0.851
22, 35, 36	Face	0.956	0.877
22, 35, 36	Place	0.957	0.902
22, 35, 36	Object	0.956	0.861
22, 35, 36, 38, 39, 40, 41	Face	0.974	0.911
22, 35, 36, 38, 39, 40, 41	Place	0.974	0.904
22, 35, 36, 38, 39, 40, 41	Object	0.973	0.889
22, 23, 24, 35, 36, 38, 39, 40, 41	Face	0.981	0.921
22, 23, 24, 35, 36, 38, 39, 40, 41	Place	0.980	0.907
22, 23, 24, 35, 36, 38, 39, 40, 41	Object	0.981	0.894

Table 4.2: ROI Labels

ROI	Label
22	Angular Gyrus
23	Superior Lateral Occipital Cortex
24	Inferior Lateral Occipital Cortex
35	Anterior Parahippocampus
36	Posterior Parahippocampus
38	Anterior Division of Temporal Fusiform Cortex
39	Posterior Division of Temporal Fusiform Cortex
40	Temporal Occipital Fusiform Cortex
41	Occipital Fusiform Cortex

Table 4.3: Binary Classification of Category vs. Rest (Across Runs)

ROIs	Category	Train Accuracy	Test Accuracy
22	Face	0.772	0.649
22	Place	0.810	0.733
22	Object	0.801	0.692
23, 24	Face	0.842	0.789
23, 24	Place	0.875	0.829
23, 24	Object	0.813	0.757
35, 36	Face	0.832	0.743
35, 36	Place	0.892	0.832
35, 36	Object	0.786	0.733
22, 35, 36	Face	0.949	0.877
22, 35, 36	Place	0.942	0.908
22, 35, 36	Object	0.947	0.841
22, 35, 36, 38, 39, 40, 41	Face	0.974	0.933
22, 35, 36, 38, 39, 40, 41	Place	0.964	0.932
22, 35, 36, 38, 39, 40, 41	Object	0.966	0.872
22, 23, 24, 35, 36, 38, 39, 40, 41	Face	0.983	0.938
22, 23, 24, 35, 36, 38, 39, 40, 41	Place	0.976	0.937
22, 23, 24, 35, 36, 38, 39, 40, 41	Object	0.971	0.897

Table 4.4: Four-Way Classification of Category (incl. Rest) Within Runs

ROIs	Train Accuracy	Test Accuracy
22	0.628	0.501
23, 24	0.638	0.548
35, 36	0.536	0.475
22, 35, 36	0.766	0.561
22, 35, 36, 38, 39, 40, 41	0.850	0.596
22, 23, 24, 35, 36, 38, 39, 40, 41	0.885	0.589

Furthermore, for the most inclusive ROIs, there is roughly a 20% gap between train and test accuracy, suggesting a decent amount of overfitting is happening in our models. It is not clear what the cause of the overfitting is, especially since it should be reduced in the across-run condition. This is likely hurting out-of-sample accuracy, and future work should aim to minimize this gap. That being said, our results are consistent with findings in the literature, suggesting our pipeline is set up properly.

Tables 4.6 and 4.7 similarly report classification within- and across-runs, but for three-way classification, *i.e.* excluding rest and only predicting the three categories of face, place, and object. Like the previous results, performance tends to be maximized across-run with the most inclusive subset of ROIs. Peak performance reaches 55%, with chance being 33%. Rest is fundamentally different than activation of the three categories, so it is perhaps not surprising the degree to which classifier accuracy drops off. Like the previous results, there is a large gap between train and test accuracy, suggesting that there may be an opportunity to reduce overfitting and increase performance.

Lastly, Tables 4.8 and 4.9 report classifier results when pitting two categories against each other (*e.g.* face

Table 4.5: Four-Way Classification of Category (incl. Rest) Across Runs

ROIs	Train Accuracy	Test Accuracy
22	0.565	0.407
23, 24	0.627	0.502
35, 36	0.591	0.447
22, 35, 36	0.739	0.567
22, 35, 36, 38, 39, 40, 41	0.835	0.635
22, 23, 24, 35, 36, 38, 39, 40, 41	0.857	0.663

Table 4.6: Three-Way Classification of Category Within Runs

ROIs	Train Accuracy	Test Accuracy
22	0.603	0.431
23, 24	0.453	0.338
35, 36	0.347	0.296
22, 35, 36	0.688	0.426
22, 35, 36, 38, 39, 40, 41	0.803	0.452
22, 23, 24, 35, 36, 38, 39, 40, 41	0.855	0.467

Table 4.7: Three-Way Classification of Category Across Runs

ROIs	Train Accuracy	Test Accuracy
22	0.583	0.436
23, 24	0.653	0.466
35, 36	0.632	0.387
22, 35, 36	0.656	0.449
22, 35, 36, 38, 39, 40, 41	0.781	0.521
22, 23, 24, 35, 36, 38, 39, 40, 41	0.811	0.555

Table 4.8: Binary Category Classification Within Runs

ROIs	Category 1	Category 2	Train Accuracy	Test Accuracy
22	Face	Place	0.732	0.614
22	Face	Object	0.767	0.645
22	Place	Object	0.730	0.589
23, 24	Face	Place	0.715	0.591
23, 24	Face	Object	0.700	0.573
23, 24	Place	Object	0.574	0.460
35, 36	Face	Place	0.520	0.455
35, 36	Face	Object	0.583	0.462
35, 36	Place	Object	0.511	0.475
22, 35, 36	Face	Place	0.808	0.596
22, 35, 36	Face	Object	0.814	0.614
22, 35, 36	Place	Object	0.778	0.547
22, 35, 36, 38, 39, 40, 41	Face	Place	0.909	0.603
22, 35, 36, 38, 39, 40, 41	Face	Object	0.888	0.649
22, 35, 36, 38, 39, 40, 41	Place	Object	0.837	0.554
22, 23, 24, 35, 36, 38, 39, 40, 41	Face	Place	0.946	0.652
22, 23, 24, 35, 36, 38, 39, 40, 41	Face	Object	0.925	0.641
22, 23, 24, 35, 36, 38, 39, 40, 41	Place	Object	0.873	0.540

Table 4.9: Binary Category Classification Across Runs

ROIs	Category 1	Category 2	Train Accuracy	Test Accuracy
22	Face	Place	0.705	0.594
22	Face	Object	0.761	0.636
22	Place	Object	0.712	0.609
23, 24	Face	Place	0.772	0.605
23, 24	Face	Object	0.821	0.658
23, 24	Place	Object	0.712	0.576
35, 36	Face	Place	0.708	0.513
35, 36	Face	Object	0.799	0.531
35, 36	Place	Object	0.692	0.558
22, 35, 36	Face	Place	0.752	0.605
22, 35, 36	Face	Object	0.801	0.643
22, 35, 36	Place	Object	0.772	0.592
22, 35, 36, 38, 39, 40, 41	Face	Place	0.873	0.701
22, 35, 36, 38, 39, 40, 41	Face	Object	0.884	0.679
22, 35, 36, 38, 39, 40, 41	Place	Object	0.828	0.625
22, 23, 24, 35, 36, 38, 39, 40, 41	Face	Place	0.915	0.759
22, 23, 24, 35, 36, 38, 39, 40, 41	Face	Object	0.895	0.688
22, 23, 24, 35, 36, 38, 39, 40, 41	Place	Object	0.844	0.634

versus object). This gives a sense of the pairwise similarity between two categories and helps isolate where confusion in the more complex models may be happening. Place and face activation seem to be the most separable, consistent with contrast results suggesting they are the most specified. The object comparisons are less accurate, perhaps due to the broader cognitive control network activation in our object contrasts. This suggests again the need to isolate the LOC in subsequent analyses.

Tables A.1 to A.6 in the Appendix report classifier results averaged across all subjects, with mean accuracy and SEM reported. For each subject, a 3 to 5 TR offset was used to map between input and output, with the selected offset being the one that optimized predictive accuracy across all ROI's.

Future work needs to aim to improve the classification accuracy of the decoding models. State-of-the-art models can reach roughly 90%, and while this is likely implausible for the current data, some improvements should be possible. The difference between train and test accuracy suggests there is some overfitting going on. One of the major ways to change overfitting would be to change the model itself. Currently, we use a regularized support vector machine with a linear kernel, which was selected because we thought the model was sufficiently simple (and thus not as susceptible to overfitting). Other models, *e.g.* logistic regression [173], may function as a better alternative, though the computational demands will be greatly increased.

Other helpful techniques may aim to augment the data itself. For example, rather than train on adjacent BOLD images which will have a significant amount of overlap, a small set of 'motifs' can be averaged from the data and used to train the model. Currently, each run has four blocks each of the three categories. Perhaps, we could generate four motifs each for twelve motifs (and a total of twenty-four across two runs). While this

is a small training set, perhaps a sufficiently simple model (*e.g.* a support vector machine with a linear kernel) would lead to greater accuracy.

A second way to modify the data is to add manual convolution. Currently, we train on and predict each TR independently. However, we know that neighboring TR's are likely to be related to one another. Thus, embedding manual convolution into our classifiers would involve giving data about neighboring TRs and allow us to weight them in a manner that optimizes performance.

In summary, while these results are promising, there is work to be done on improvement. Once the classifiers are sufficiently accurate, we can decode the 'rest' periods between every trial to get a proxy for what categories are being replayed (or reactivated). For every individual, we have trial-by-trial predictions as to what categories are being replayed, and we can evaluate the extent to which these predictions correspond to the neural decoding, and subsequently, the extent to which we can explain individual differences in remote updating.

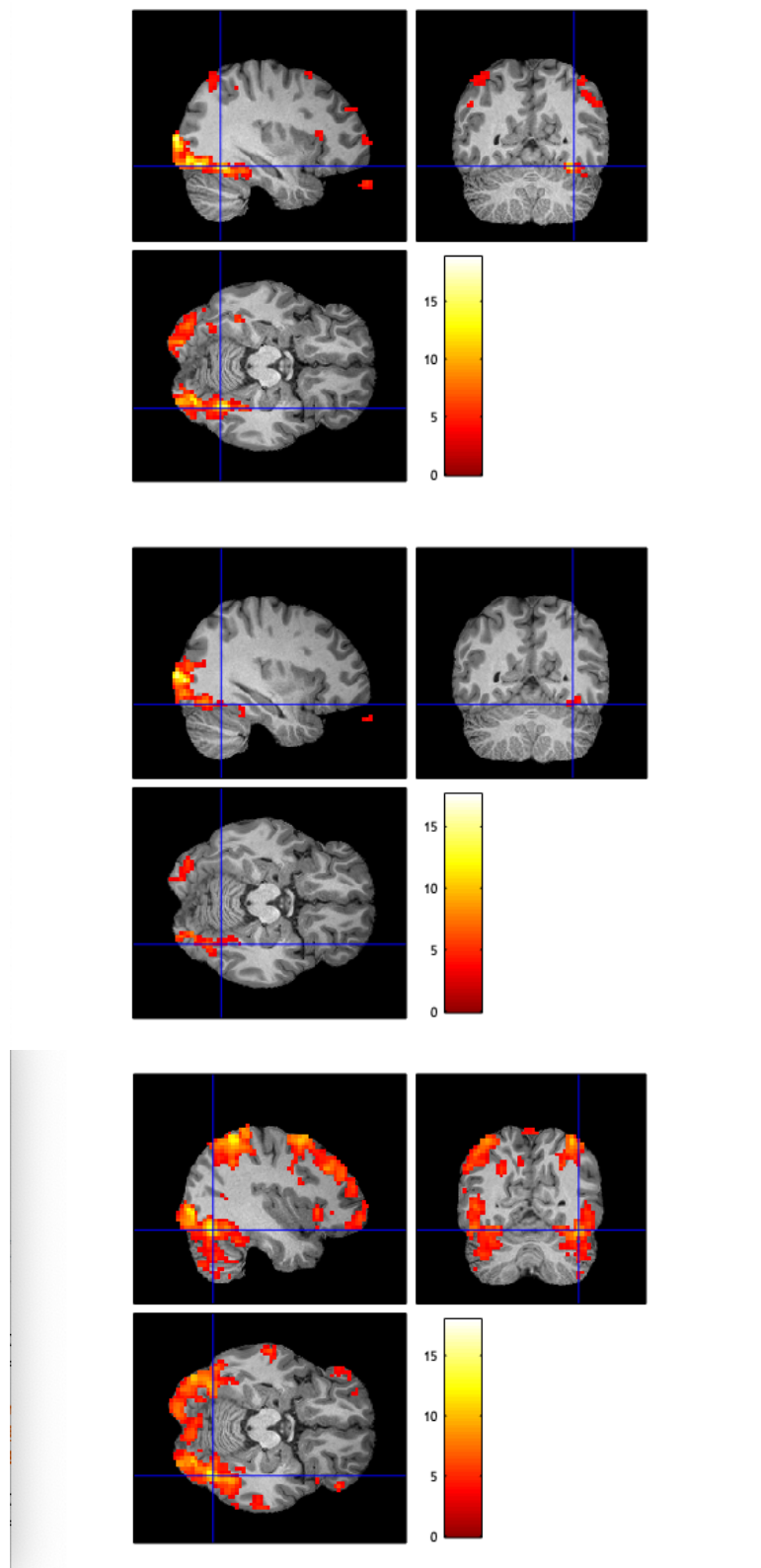


Figure 4.3: Face, place, and object contrasts for one individual subject in MNI space. Cluster minimum of 10 voxels at an uncorrected $p < 0.001$.

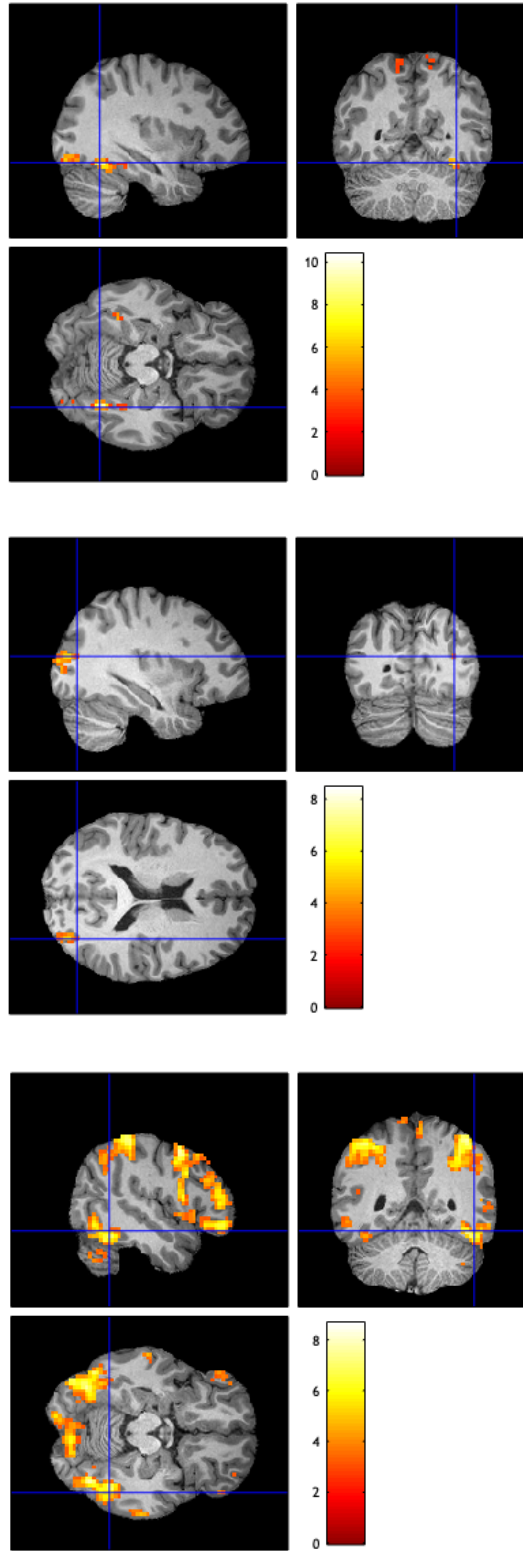


Figure 4.4: Face vs. all, place vs. all, and object vs. all contrasts for one individual subject in MNI space. Cluster minimum of 10 voxels at an uncorrected $p < 0.001$.

4.3 Task Contrasts

We next compute contrasts on different event types during the behavioral task. Participants performed 200 trials, and we divided this session over two different scanning runs, allowing participants to take a short break between scans. These two scanning sequences were parameterized in the same way as the two previous localizer scans.

Similar to the localizer contrasts above, we perform first-level analyses on each subject by combining both runs and we then perform second-level analyses over all the analyzed subjects. All analyses are in MNI space and show voxel (minimum cluster size of 10) activation at an uncorrected $p < 0.001$. In addition to standard motion correct, five specific variables of interest were regressed in the design matrix: button press, (no) reward, reward, bonus reward, regular trial, and bonus trials.

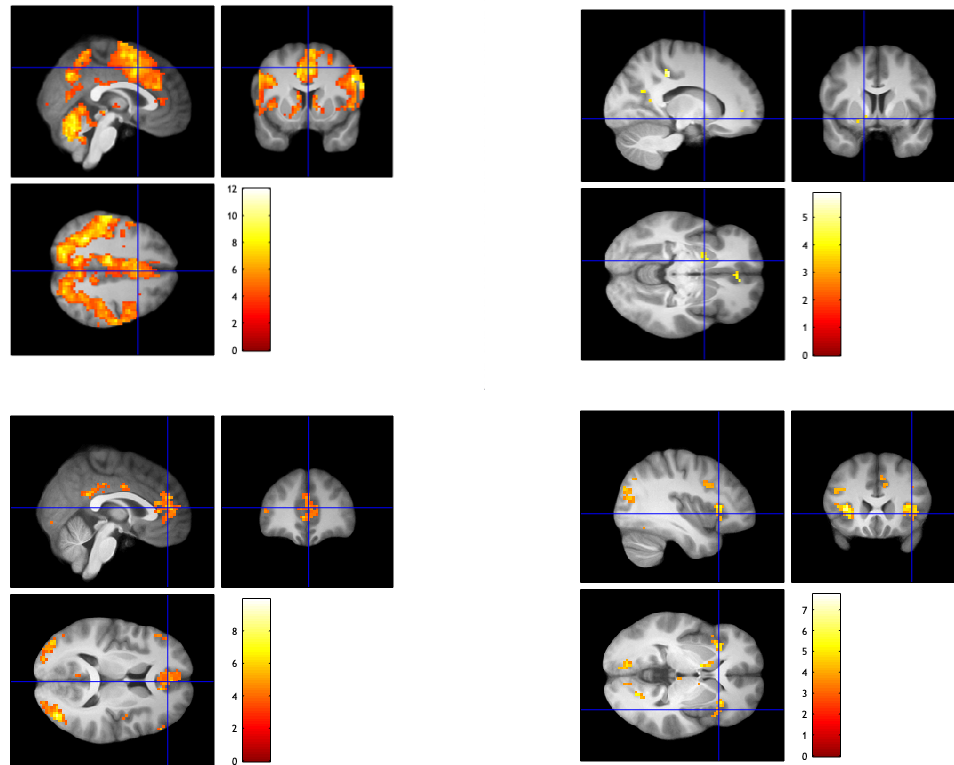


Figure 4.5: Four different contrasts during the two task runs across subjects. Cluster minimum of 10 voxels with an uncorrected p -value of 0.001. Anatomicals are in MNI space. (Top Left) Button Press. (Top Right) Reward minus No Reward. (Bottom Left) Bonus Reward minus No Reward. (Bottom Right) Bonus Trials minus Regular Trials.

Figure 4.5 plots four relevant contrasts. The first, in the top left, plots contrasts corresponding to when participants pressed a button. As expected, we see widespread activation across the premotor and motor cortices, confirming that our scanning and time tracking are aligned. Furthermore, the contrast demonstrates

that the left side is preferentially activated, presumably corresponding to our participants being right-handed (and thus pressing the button with their right hand). This initial contrast acts as a sanity check to ensure our pipeline is intact.

Next, in the top right, we plot the differences between rewarded trials and not rewarded trials. The difference between these two trials gives a proxy for the magnitude of reward prediction error. Reinforcement learning models were famously introduced in neuroscience as a model that connects dopamine activation in rodents to the computational imperative of prediction error, and subsequent fMRI work has seen similar results in humans [48, 55]. Our results in the top right plot align with this work, where we see activation in a cluster corresponding to striatum, which is the location of dopaminergic neurons that are hypothesized to fire in proportion to reward prediction errors. Reward prediction errors are much more temporally sensitive than button pressing, and thus our results further suggest our preprocessing and alignment pipeline is valid.

The bottom left plots imaging contrasts for the differences between rewards during bonus trials and no rewarded trials. We expected to see even greater striatal activation than in the previous plot as prediction errors should be greater, but this is not the case. However, there does seem to be systematic activation in visual cortex, posterior cingulate, and anterior cingulate. Visual cortex activation likely arises from the colorful effects of reward (points are awarded in a rainbow-colored text). Cingulate activation, especially anterior cingulate, is especially intriguing because this has been implicated as a key metacognitive decision-making system in rational models of control [190]. Our theory suggests a similar computation is being done when determining whether to replay or not. As a result, we have offered the anterior cingulate as a potential region of interest to where $EV B_C$ calculations are occurring.

Lastly, the bottom right plots the contrast between bonus trials and regular trials. In addition to the visual cortical activation (which is expected due to the colorful stimuli), we see systematic bilateral insular activation. However, we are relatively agnostic in interpreting this effect.

4.4 Default Mode Network

The default mode network (DMN) refers to a set of brain regions (primarily medial prefrontal cortex and posterior cingulate) that are systematically activated during periods of wakeful rest. Despite these consistent findings, many scientists are often puzzled by its function. Recent work has suggested there is a coupling between replay and default mode [87, 100, 154], and this coupling may coalesce well with our proposal that the phenomenology of cognitive fatigue is related to the value of hippocampal replay.

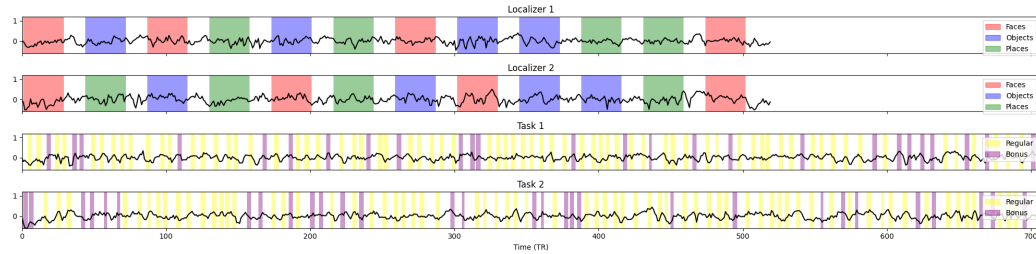


Figure 4.6: Z-scored Default Mode Network (DMN) Activation across four scanning runs for a sample subject. The first two runs are localizer runs, and shaded portions correspond to category stimuli. The second two runs are the task runs, and shaded portions correspond to task stimuli.

Figure 4.6 plots Z-scored default mode activation for a sample subject across the four scanning runs (two localizer runs, two task runs). To compute this activation, we align the Schaefer atlas [177] with the subject’s native space. We isolate regions of interest (ROIs) that correspond to the pre-specified default mode network. We then regress out motion regressors, compute the mean activation, and Z-score this activation to see how it changes through the course of a run.

First, as a sanity check, we look at the localizer runs and test whether default mode activation is higher during off periods compared to image stimuli. Figure 4.7 plots two comparisons. First, on the left, we plot default mode activation during rest versus all the localizer stimuli. A t-test based on mean activation find DMN activity during rest to be statistically higher ($p < 0.05$) compared to the periods in which the localizer stimuli are present on the screen. Analyses use an offset of four TRs between stimulus and BOLD activity as this number optimized the difference between the two conditions.

The right of Figure 4.7 plots a finer-grained comparison, comparing DMN activity in the localizer rest period compared to each image category (face, place, object) separately. For our sample subject, we see that activation is not significantly different between rest and face, and rest and object, but the difference is significant between rest and place. These results are consistent with our previous localizer results, which demonstrated that posterior cingulate (a region implicated in the default mode network) activation is higher during face stimuli and that the cognitive control area is activated during object activation (presumably due to our attention task), whereas place activity is well-localized to the PPA.

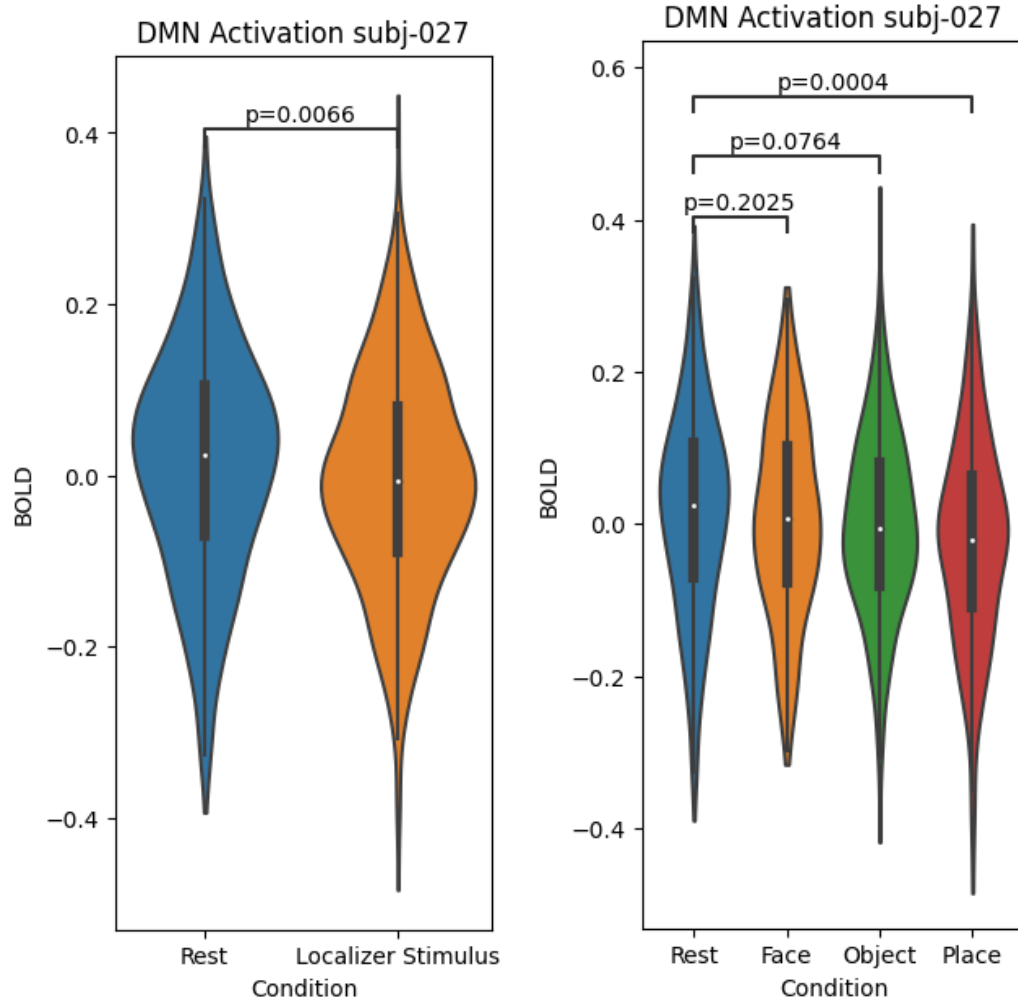


Figure 4.7: Default mode network (DMN) activation between periods of no stimulus versus stimulus during the two localizer runs.

Figure 4.8 plots similar results, but this time averaging across all subjects. For each subject, we take the mean Z-scored activation for each condition (rest, face, object, place), and compare mean activations across subjects. Figure 4.8 demonstrates that DMN activation is significantly higher at rest compared to stimulus at the $p < 0.0001$ and that this relationship holds when separating out based on stimulus conditions. Like the previous analyzes, these effects are optimized when offsets are between 3 - 5 TRs (which we individually optimized for each subject).

Next, we evaluate the extent to which variance in default mode activation relates to the model-based effect in each participant. One of our hypotheses is that default mode activation is tracking replay activity, and if replay is captured by the model-based behavior of a participant, there may be a relationship between DMN activation and model-based behavior that captures individual differences.

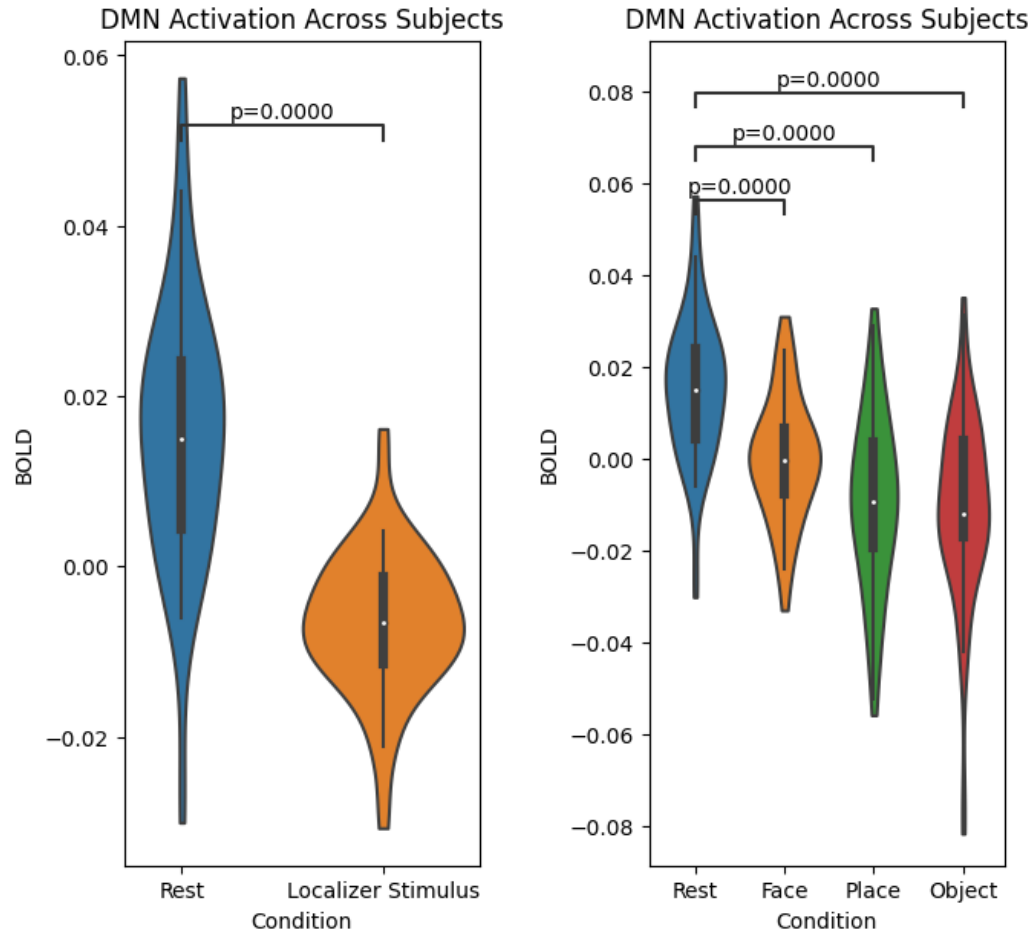


Figure 4.8: Across-subject default mode network (DMN) activation between periods of no stimulus versus stimulus during the two localizer runs.

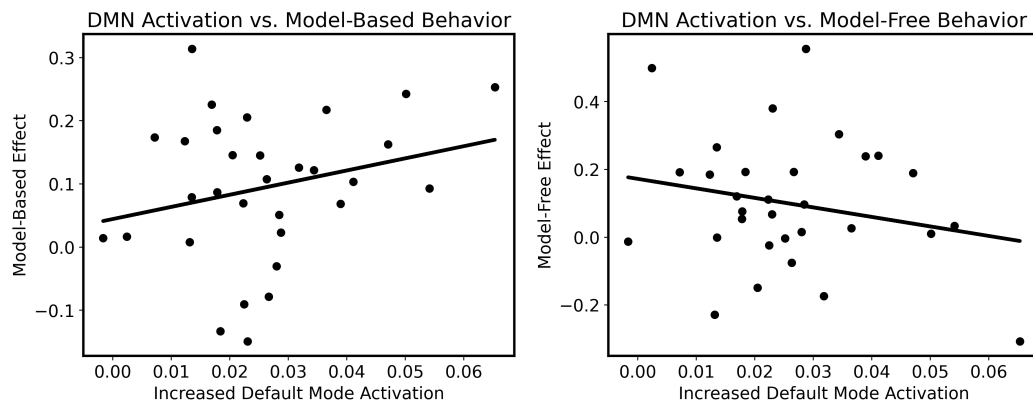


Figure 4.9: DMN activation versus model-based and model-free behavior across all subjects

Figure 4.9 reports the results of these analyses, with the left reporting the correlation with model-based behavior and the right reporting the correlation with model-free behavior. We calculate the model-based and model-free effects from the win-stay, lose-shift behavior, and thus we all include all subjects (not just the

‘good’ subjects since we want to explain individual differences). There is a positive correlation between DMN activation and model-based behavior ($r = 0.25, p = 0.17$), though this effect is not significant at the $p < 0.05$ level. The effect increases when ‘bad’ participants are excluded, but the effect remains insignificant. This gives weak evidence that there may be a relationship between DMN activation and model-based behavior, and perhaps a higher-powered study could identify a significant relationship. DMN activation and model-free behavior are negatively correlated ($r = -0.22, p = 0.24$), though the effect is not statistically significant.

Future, more precise work should break up DMN activation into three hypothesized subregions as well as continue masking out pre-defined ROIs for the localizer classifiers.

4.5 Discussion

This chapter demonstrated a proof-of-concept that the multi-echo neuroimaging pipeline was collected properly and was adequately aligned with the task. Sanity contrasts such as button press activation in motor cortex and reward prediction errors in the striatum were consistent with past reinforcement learning neuroimaging studies, and the visual category (face, place, object) contrasts generally recapitulated with past literature.

Future work will use the localizer classifiers to make predictions on what events are being replayed during the task. Three hypotheses are of interest here. In the first, we will determine whether replay and default mode activation track one another. Specifically, we can use the four-way classifiers to determine whether the participant is not replaying (prediction of rest) versus whether they are replaying one of the three categories. Either could be correlated with default mode activation; the former suggesting a relationship between ‘off-task’ and DMN activation, and the latter suggesting DMN activation facilitates replay. While we did not see a strong relationship between model-based updating and DMN activation, an alternative approach using localizer classifiers may shed some light in either direction.

Second, we can evaluate the extent to which replay follows the *EVB* model [130], as done in [122]. We manipulate both the *Need* and *Gain* terms, and thus we can make trial-by-trial predictions for what a participant is replaying and then see whether the localizer classifier agrees.

Last, we can determine whether the strength of replay as inferred by both the logistic regression models and the hierarchical Bayesian models are correlated with the strength of replay as inferred by the localizer classifier. Specifically, we would predict that participants who have more model-based updating should have stronger predictions of replay (or reactivation) during their trials.

Chapter 5

Conclusion

Over time, our ability to exert cognitive control seemingly diminishes, and this results in the phenomenological state of cognitive fatigue and the corresponding decrements in task performance. Yet, while cognitive fatigue has been studied for over a hundred years, its origins still remain a mystery. The most influential line of work (ego depletion) suggested fatigue to reflect the depletion of a metabolic resource, and while this idea largely fell out of favor, it has given rise to a widely agreed upon insight: the value of rest. We have built a line of research that extends previous work on hippocampal replay, a widely studied neural mechanism active during ‘rest.’ We proposed that cognitive control mitigates the use of replay, and we demonstrated that this specification reproduces the behavioral effects of cognitive fatigue [5]. We then used this model to construct a taxonomy of the related phenomena of fatigue, boredom, and mind-wandering; we demonstrated their computational similarities while also offering dissociable functions, paving the way for future research to study them independently [34]. Furthermore, the link between fatigue and replay lets us build upon a rich tradition in the reinforcement learning community to use neuroimaging as a tool by which to validate trial-to-trial behavioral variation [3]. Here, we are seeking to confirm that control is a mechanism that modulates the use of replay, using both neural decoding and computational modeling of behavior. Future work will determine how agents *perceive* time and how does this perception affect judgments of effort? Colloquially, time flies quickly when you’re having fun and drags when you’re bored. The ‘contextual change’ hypothesis [27, 26] explains people’s time perception judgments as a function of the number of context shifts they have experienced. We can formalize this idea by building computational models of time perception. Specifically, we can use Bayesian nonparametric inference [73] to make precise predictions on the number of context switches and thus the perceived time duration for any given task. We can then integrate these behavioral and computational results with related neuroimaging work in ‘event segmentation theory’ [223, 68] to construct

a unified behavioral, computational, and neural theory of time perception.

5.1 Cost of Time

This thesis leverages an increasingly popular insight: that time is *costly*. Agents have limited resources, and with limited time, there is a need for prioritization. This thesis explored this concept through the phenomenology of cognitive fatigue and boredom. Below are two other lines of work that leverage the cost of time: (1) the scarcity of researchers’ resources and the opportunity to evaluate thousands of hypotheses simultaneously and (2) the rationality of using heuristics and its philosophical relevance.

5.1.1 Big Data and Machine Learning for Theoretical Progress.

With (tens of) thousands of dollars and off-the-shelf machine learning algorithms, (behavioral) scientists can train predictive models that outperform symbolic theories. Is theory dead? We argue no. Theoretical explanations are imperative to the scientific endeavor: they influence how we think about problems and enable us to generalize findings from one domain to another. Thus, we have started to adapt the scientific method to incorporate machine learning algorithms so that scientists can jointly maximize both the predictive and explanatory power of their theories. We proposed a methodology called ‘Scientific Regret Minimization’ that helps scientists use the predictive power of machine learning to specify shortcomings in extant theories and identify novel empirical phenomena [7]. We demonstrated this method on the Moral Machine dataset [17] in order to (i) build a complex, but interpretable computational model of moral judgment and (ii) use this model to identify three increasingly complex and novel moral behaviors. We then applied this methodology to two classic decision-making problems: [164] and intertemporal choice [6]. In risky choice, we found that an unconstrained machine learning model can outperform over fifty models in the literature, and yet can be simplified into a mixture of two previous ideas: an expected utility component that accounts for the vast amount of choice problems and a prospect theory component that accounts for classic biases when facing complex problems. This new theory encapsulates past work while specifying the new question for future research: what problem characteristics determine an individual’s strategy? Our work in intertemporal choice demonstrates that unconstrained neural networks outperform classic discounting functions, and we are currently building a theoretically-driven and interpretable model that can match the neural network’s performance and accommodate individual variation. Future work will use similar tools to construct ‘basis sets’ of human decision-making biases. Psychologists and behavioral economists have catalogued a significant number of biases [206, 52], and we seek to understand the extent to which these biases are independent of one another. We can collect large-scale datasets that capture the different decision-making biases and then fit

machine learning algorithms to subsets of these biases in order to which models reproduce the left-out biases. These metrics can be analyzed to create a ‘basis set’ of decision-making biases, which we will then validate through fine-grained individual behavioral studies.

5.1.2 Integrating Theoretical and Practical Rationality in Philosophy

Experimental psychology has largely invalidated the *homo economicus* assumption which posited that humans obey the classical axioms of rationality. While early research suggested these results indicate human irrationality, recent work views them as reflecting the rational use of limited, cognitive resources [193, 194, 118, 119]. This perspective has started to offer a new standard of rationality. We are using this perspective to reconcile one of the biggest tensions of the last few centuries in moral philosophy: consequentialism versus deontology. Rule consequentialism [137, 83, 90] offered an exciting compromise between these two traditions, proposing agents ought to abide by deontic principles (i.e. rules) which themselves are justified by the consequentialist calculus, however, it largely fell out of favor in the mid-twentieth century due to its perceived instability [176, 13]. We have resurrected this idea, leveraging the past few decades of experimental data and computational modeling to (i) provide a normative moral theory that reconciles long-held tensions and (ii) provide a new descriptive theory in which to understand human behavior [4].

Future work along this line will expand the scope of bounded rationality to other philosophical problems. As one illustrative example, consider scientific virtues. Philosophers of science suggest our scientific hypotheses should have values such as simplicity and unification, even if these values don’t directly correspond to the evidentiary confirmation [112]. Bounded rationality is a framework that can rationalize these virtues, which would enable us to integrate epistemic and pragmatic approaches to scientific hypotheses.

5.2 Concluding Remarks

We have seen cognitive science come of age in the last few decades. Technologically, we now have access to a plethora of tools that enable us to abstract a psychological problem into its most general form, construct a set of possible solutions, and then validate a subset of them through neural and biological data. This thesis integrated work in psychology, neuroscience, computer science, and philosophy. Here, we aimed to formalize an elusive phenomenological construct in the language of computation, and we then designed a behavioral and neuroimaging study to tie empirical data to computational imperatives.

Bibliography

- [1] N. Afzalian and R. Rajimehr. Spatially adjacent regions in posterior cingulate cortex represent familiar faces at different levels of complexity. *Journal of Neuroscience*, 41(47):9807–9826, 2021.
- [2] S. Agnoli, M. Vanucci, C. Pelagatti, and G. E. Corazza. Exploring the link between mind wandering, mindfulness, and creativity: A multidimensional approach. *Creativity Research Journal*, 30(1):41–53, 2018.
- [3] M. Agrawal, S. Allan, I. D. Campbell, M. G. Mattar, J. D. Cohen, and N. D. Daw. Measuring opportunity costs and hippocampal experiential activation. in progress.
- [4] M. Agrawal and D. Danks. Maxim consequentialist for bounded agents. under review.
- [5] M. Agrawal, M. G. Mattar, J. D. Cohen, and N. D. Daw. The temporal dynamics of opportunity costs: A normative account of cognitive fatigue and boredom. *Psychological review*, 129(3):564, 2022.
- [6] M. Agrawal, J. C. Peterson, J. D. Cohen, and T. L. Griffiths. Stress, intertemporal choice, and mitigation behavior during the covid-19 pandemic. *Journal of Experimental Psychology: General*, 2023.
- [7] M. Agrawal, J. C. Peterson, and T. L. Griffiths. Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences*, 117(16):8825–8835, 2020.
- [8] J. R. Anderson. Skill acquisition: Compilation of weak-method problem situations. *Psychological review*, 94(2):192, 1987.
- [9] P. W. Andrews and J. A. Thomson Jr. The bright side of being blue: depression as an adaptation for analyzing complex problems. *Psychological review*, 116(3):620, 2009.
- [10] J. Y. Angela and P. Dayan. Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692, 2005.

- [11] J. W. Antony, T. H. Hartshorne, K. Pomeroy, T. M. Gureckis, U. Hasson, S. D. McDougale, and K. A. Norman. Behavioral, physiological, and neural signatures of surprise during naturalistic sports viewing. *Neuron*, 109(2):377–390, 2021.
- [12] T. Arai. *Mental fatigue*. Number 54. Teachers College, Columbia University, 1912.
- [13] R. Arneson. Sophisticated rule consequentialism: some simple objections. *Philosophical issues*, 15:235–251, 2005.
- [14] J. Ashburner. Spm: a history. *Neuroimage*, 62(2):791–800, 2012.
- [15] G. Aston-Jones and J. D. Cohen. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, 28:403–450, 2005.
- [16] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.
- [17] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan. The moral machine experiment. *Nature*, 563:59–64, 2018.
- [18] C. Baldassano, D. M. Beck, and L. Fei-Fei. Differential connectivity within the parahippocampal place area. *Neuroimage*, 75:228–237, 2013.
- [19] S. H. Bartley and E. Chute. Fatigue and impairment in man. 1947.
- [20] R. F. Baumeister, E. Bratslavsky, M. Muraven, and D. M. Tice. Ego depletion: Is the active self a limited resource? *Journal of personality and social psychology*, 74(5):1252, 1998.
- [21] R. F. Baumeister and K. D. Vohs. Self-regulation, ego depletion, and motivation. *Social and personality psychology compass*, 1(1):115–128, 2007.
- [22] S. W. Bench and H. C. Lench. On the function of boredom. *Behavioral sciences*, 3(3):459–472, 2013.
- [23] B. O. Bergum and D. J. Lehr. Vigilance performance as a function of interpolated rest. *Journal of Applied Psychology*, 46(6):425, 1962.
- [24] D. A. Berry and B. Fristedt. Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). London: Chapman and Hall, 5(71-87):7–7, 1985.
- [25] B. Blain, G. Hollard, and M. Pessiglione. Neural mechanisms underlying the impact of daylong cognitive work on economic decisions. *Proceedings of the National Academy of Sciences*, 113(25):6967–6972, 2016.

- [26] R. A. Block. Temporal judgments and contextual change. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(6):530, 1982.
- [27] R. A. Block and M. A. Reed. Remembered duration: Evidence for a contextual-change hypothesis. *Journal of Experimental psychology: Human Learning and memory*, 4(6):656, 1978.
- [28] K. C. Bloom and T. J. Shuell. Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *The Journal of Educational Research*, 74(4):245–248, 1981.
- [29] T. S. Braver, D. M. Barch, and J. D. Cohen. Cognition and control in schizophrenia: a computational model of dopamine and prefrontal function. *Biological psychiatry*, 46(3):312–328, 1999.
- [30] P.-C. Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28, 2017.
- [31] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.
- [32] M. F. Carr, S. P. Jadhav, and L. M. Frank. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nature neuroscience*, 14(2):147, 2011.
- [33] E. C. Carter, L. M. Kofler, D. E. Forster, and M. E. McCullough. A series of meta-analytic tests of the depletion effect: self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, 144(4):796, 2015.
- [34] I. Christian and M. Agrawal. A computational model of unintentional mind wandering in focused attention meditation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022.
- [35] K. Christoff, Z. C. Irving, K. C. Fox, R. N. Spreng, and J. R. Andrews-Hanna. Mind-wandering as spontaneous thought: a dynamic framework. *Nature Reviews Neuroscience*, 17(11):718, 2016.
- [36] J. J. Clarkson, E. R. Hirt, L. Jia, and M. B. Alexander. When perception is more than reality: the effects of perceived versus actual resource depletion on self-regulatory behavior. *Journal of personality and social psychology*, 98(1):29, 2010.
- [37] G. Cohen and D. M. Burke. Memory for proper names: A review. *Memory*, 1(4):249–263, 1993.
- [38] J. D. Cohen, K. Dunbar, and J. L. McClelland. On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychological review*, 97(3):332, 1990.

- [39] J. D. Cohen, S. M. McClure, and A. J. Yu. Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):933–942, 2007.
- [40] J. D. Cohen and D. Servan-Schreiber. Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychological review*, 99(1):45, 1992.
- [41] M. W. Cole and W. Schneider. The cognitive control network: Integrated cortical regions with dissociable functions. *Neuroimage*, 37(1):343–360, 2007.
- [42] L. H. Corbit, S. B. Ostlund, and B. W. Balleine. Sensitivity to instrumental contingency degradation is mediated by the entorhinal cortex and its efferents via the dorsal hippocampus. *Journal of Neuroscience*, 22(24):10976–10984, 2002.
- [43] M. Csikszentmihalyi. *Creativity: The work and lives of 91 eminent people*. HarperCollins, 1996.
- [44] M. Csikszentmihalyi. Flow and the psychology of discovery and invention. *HarperPerennial, New York*, 39, 1997.
- [45] W. Dabney, Z. Kurth-Nelson, N. Uchida, C. K. Starkweather, D. Hassabis, R. Munos, and M. Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, pages 1–5, 2020.
- [46] I. Dasgupta, K. A. Smith, E. Schulz, J. B. Tenenbaum, and S. J. Gershman. Learning to act by integrating mental simulations and physical experiments. *BioRxiv*, page 321497, 2018.
- [47] T. J. Davidson, F. Kloosterman, and M. A. Wilson. Hippocampal replay of extended experience. *Neuron*, 63(4):497–507, 2009.
- [48] N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, and R. J. Dolan. Model-based influences on humans’ choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.
- [49] N. D. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704, 2005.
- [50] N. D. Daw, J. P. O’doherly, P. Dayan, B. Seymour, and R. J. Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879, 2006.
- [51] P. Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

- [52] B. De Martino, D. Kumaran, B. Seymour, and R. J. Dolan. Frames, biases, and rational decision-making in the human brain. *Science*, 313(5787):684–687, 2006.
- [53] K. Diba and G. Buzsáki. Forward and reverse hippocampal place-cell sequences during ripples. *Nature neuroscience*, 10(10):1241, 2007.
- [54] R. Dodge. The laws of relative fatigue. *Psychological Review*, 24(2):89, 1917.
- [55] B. B. Doll, K. D. Duncan, D. A. Simon, D. Shohamy, and N. D. Daw. Model-based choices involve prospective neural activity. *Nature neuroscience*, 18(5):767–772, 2015.
- [56] J. J. Donovan and D. J. Radosevich. A meta-analytic review of the distribution of practice effect: Now you see it, now you don’t. *Journal of Applied Psychology*, 84(5):795, 1999.
- [57] J. Dora, M. van Hooff, S. Geurts, M. Kompier, and E. Bijleveld. The effect of opportunity costs on mental fatigue in labor/leisure tradeoffs. 2019.
- [58] R. Dubey and T. Griffiths. Reconciling novelty and complexity through a rational analysis of curiosity. 2019.
- [59] E. DuPre, T. Salo, Z. Ahmed, P. A. Bandettini, K. L. Bottenhorn, C. Caballero-Gaudes, L. T. Dowdle, J. Gonzalez-Castillo, S. Heunis, P. Kundu, et al. Te-dependent analysis of multi-echo fmri with* tedana. *Journal of Open Source Software*, 6(66):3669, 2021.
- [60] E. Eldar, G. Lièvre, P. Dayan, and R. J. Dolan. The roles of online and offline replay in planning. *BioRxiv*, 2020.
- [61] R. Epstein and N. Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.
- [62] O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.
- [63] C. D. Fisher. Boredom: Construct, causes and consequences. Technical report, 1987.
- [64] D. J. Foster and M. A. Wilson. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440(7084):680, 2006.
- [65] K. C. Fox and K. Christoff. Metacognitive facilitation of spontaneous thought processes: when metacognition helps the wandering mind find its way. In *The cognitive neuroscience of metacognition*, pages 293–319. Springer, 2014.

- [66] K. C. Fox and K. Christoff. *The Oxford handbook of spontaneous thought: Mind-wandering, creativity, and dreaming*. Oxford University Press, 2018.
- [67] K. C. Fox, S. Nijboer, E. Solomonova, G. W. Domhoff, and K. Christoff. Dreaming as mind wandering: evidence from functional neuroimaging and first-person content reports. *Frontiers in human neuroscience*, 7:412, 2013.
- [68] N. T. Franklin, K. A. Norman, C. Ranganath, J. M. Zacks, and S. J. Gershman. Structured event memory: A neuro-symbolic model of event cognition. *Psychological Review*, 127(3):327, 2020.
- [69] C. Gagne, P. Dayan, and S. J. Bishop. When planning to survive goes wrong: predicting the future and replaying the past in anxiety and ptsd. *Current Opinion in Behavioral Sciences*, 24:89–95, 2018.
- [70] A. Geana, R. Wilson, N. D. Daw, and J. D. Cohen. Boredom, information-seeking and exploration. In *CogSci*, 2016.
- [71] P. J. Geiwitz. Structure of boredom. *Journal of personality and social psychology*, 3(5):592, 1966.
- [72] S. J. Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42, 2018.
- [73] S. J. Gershman, D. M. Blei, and Y. Niv. Context, learning, and extinction. *Psychological review*, 117(1):197, 2010.
- [74] S. J. Gershman, A. B. Markman, and A. R. Otto. Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, 143(1):182, 2014.
- [75] S. J. Gershman, C. D. Moore, M. T. Todd, K. A. Norman, and P. B. Sederberg. The successor representation and temporal context. *Neural Computation*, 24(6):1553–1568, 2012.
- [76] E. L. Gibson. Carbohydrates and mental function: feeding or impeding the brain? *Nutrition Bulletin*, 32:71–83, 2007.
- [77] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.
- [78] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum. Topics in semantic representation. *Psychological review*, 114(2):211, 2007.
- [79] K. Grill-Spector, Z. Kourtzi, and N. Kanwisher. The lateral occipital complex and its role in object recognition. *Vision research*, 41(10-11):1409–1422, 2001.

- [80] A. S. Gupta, M. A. van der Meer, D. S. Touretzky, and A. D. Redish. Hippocampal replay is not a simple function of experience. *Neuron*, 65(5):695–705, 2010.
- [81] J. S. Haager, C. Kuhbandner, and R. Pekrun. To be bored or not to be bored—how task-related boredom influences creative performance. *The Journal of Creative Behavior*, 52(4):297–304, 2018.
- [82] M. S. Hagger, N. L. Chatzisarantis, H. Alberts, C. O. Anggono, C. Batailler, A. R. Birt, R. Brand, M. J. Brandt, G. Brewer, S. Bruyneel, et al. A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4):546–573, 2016.
- [83] R. M. Hare. *Moral thinking: Its levels, method, and point*. Oxford: Clarendon Press; New York: Oxford University Press, 1981.
- [84] A. F. Healy, J. A. Kole, C. J. Buck-Gengler, and L. E. Bourne. Effects of prolonged work on data entry speed and accuracy. *Journal of Experimental Psychology: Applied*, 10(3):188, 2004.
- [85] W. S. Helton and P. N. Russell. Rest is best: The role of rest and task interruptions on vigilance. *Cognition*, 134:165–173, 2015.
- [86] W. S. Helton and P. N. Russell. Rest is still best: The role of the qualitative and quantitative load of interruptions on vigilance. *Human factors*, 59(1):91–100, 2017.
- [87] C. Higgins, Y. Liu, D. Vidaurre, Z. Kurth-Nelson, R. Dolan, T. Behrens, and M. Woolrich. Replay bursts in humans coincide with activation of the default mode and parietal alpha networks. *Neuron*, 109(5):882–893, 2021.
- [88] G. R. J. Hockey. A motivational control theory of cognitive fatigue. *Cognitive fatigue: Multidisciplinary perspectives on current research and future applications*, pages 167–87, 2011.
- [89] R. Hockey. *The psychology of fatigue: Work, effort and control*. Cambridge University Press, 2013.
- [90] B. Hooker. Rule-consequentialism. *Mind*, 99(393):67–77, 1990.
- [91] Z. L. Huxtable, M. H. White, and M. A. McCartor. A re-performance and re-interpretation of the arai experiment in mental fatigue with three subjects. *Psychological Monographs*, 59(5):1–52, 1946.
- [92] Q. J. Huys, T. V. Maia, and M. J. Frank. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, 19(3):404, 2016.
- [93] M. Inzlicht, B. J. Schmeichel, and C. N. Macrae. Why self-control seems (but may not be) limited. *Trends in cognitive sciences*, 18(3):127–133, 2014.

- [94] C. Izawa. Massed and spaced practice in paired-associate learning: List versus item distributions. *Journal of Experimental Psychology*, 89(1):10, 1971.
- [95] S. A. Jackson and M. Csikszentmihalyi. *Flow in sports*. Human Kinetics, 1999.
- [96] S. P. Jadhav, G. Rothschild, D. K. Roumis, and L. M. Frank. Coordinated excitation and inhibition of prefrontal ensembles during awake hippocampal sharp-wave ripple events. *Neuron*, 90(1):113–127, 2016.
- [97] V. Job, C. S. Dweck, and G. M. Walton. Ego depletion—is it all in your head? implicit theories about willpower affect self-regulation. *Psychological science*, 21(11):1686–1693, 2010.
- [98] V. Job, G. M. Walton, K. Bernecker, and C. S. Dweck. Beliefs about willpower determine the impact of glucose on self-control. *Proceedings of the National Academy of Sciences*, 110(37):14837–14842, 2013.
- [99] A. Johnson and A. D. Redish. Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, 27(45):12176–12189, 2007.
- [100] K. Kaefer, F. Stella, B. L. McNaughton, and F. P. Battaglia. Replay, the default mode network and the cascaded memory systems model. *Nature Reviews Neuroscience*, 23(10):628–640, 2022.
- [101] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
- [102] G. A. Kane, E. M. Vazey, R. C. Wilson, A. Shenhav, N. D. Daw, G. Aston-Jones, and J. D. Cohen. Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience*, 17(6):1073–1083, 2017.
- [103] N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997.
- [104] M. P. Karlsson and L. M. Frank. Awake replay of remote experiences in the hippocampus. *Nature neuroscience*, 12(7):913, 2009.
- [105] M. Keramati, A. Dezfouli, and P. Piray. Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS computational biology*, 7(5):e1002055, 2011.
- [106] M. Keramati, P. Smittenaar, R. J. Dolan, and P. Dayan. Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. *Proceedings of the National Academy of Sciences*, 113(45):12868–12873, 2016.

- [107] W. K. Kirchner. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, 55(4):352, 1958.
- [108] K. Kishida. Temporal change of subsidiary behavior in monotonous work. *Journal of Human Ergology*, 2(1):75–89, 1973.
- [109] W. Kool and M. Botvinick. A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology: General*, 2014.
- [110] W. Kool, F. A. Cushman, and S. J. Gershman. When does model-based control pay off? *PLoS computational biology*, 12(8):e1005090, 2016.
- [111] W. Kool, S. J. Gershman, and F. A. Cushman. Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological science*, 28(9):1321–1333, 2017.
- [112] T. S. Kuhn. *The structure of scientific revolutions*, volume 111. Chicago University of Chicago Press, 1970.
- [113] D. Kumaran, D. Hassabis, and J. L. McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534, 2016.
- [114] P. Kundu, V. Voon, P. Balchandani, M. V. Lombardo, B. A. Poser, and P. A. Bandettini. Multi-echo fmri: a review of applications in fmri denoising and analysis of bold signals. *Neuroimage*, 154:59–80, 2017.
- [115] R. Kurzban, A. Duckworth, J. W. Kable, and J. Myers. An opportunity cost model of subjective effort and task performance. *Behavioral and brain sciences*, 36(6):661–679, 2013.
- [116] A. K. Lee and M. A. Wilson. Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, 36(6):1183–1194, 2002.
- [117] S. W. Lee, S. Shimojo, and J. P. O’Doherty. Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3):687–699, 2014.
- [118] R. L. Lewis, A. Howes, and S. Singh. Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science*, 6(2):279–311, 2014.
- [119] F. Lieder and T. L. Griffiths. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43, 2020.

- [120] F. Lieder, T. L. Griffiths, and M. Hsu. Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, 125(1):1, 2018.
- [121] Y. Liu, R. J. Dolan, Z. Kurth-Nelson, and T. E. Behrens. Human replay spontaneously reorganizes experience. *Cell*, 178(3):640–652, 2019.
- [122] Y. Liu, M. G. Mattar, T. E. Behrens, N. D. Daw, and R. J. Dolan. Experience replay is associated with efficient nonlocal learning. *Science*, 372(6544):eabf1357, 2021.
- [123] M. M. Lorist, M. A. Boksem, and K. R. Ridderinkhof. Impaired cognitive control and reduced cingulate activity during mental fatigue. *Cognitive Brain Research*, 24(2):199–205, 2005.
- [124] K. Louie and M. A. Wilson. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, 29(1):145–156, 2001.
- [125] C. J. Lynch, I. Elbau, and C. Liston. Improving precision functional mapping routines with multi-echo fmri. *Current opinion in behavioral sciences*, 40:113–119, 2021.
- [126] J. F. Mackworth. Vigilance, arousal, and habituation. *Psychological review*, 75(4):308, 1968.
- [127] N. H. Mackworth. The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, 1(1):6–21, 1948.
- [128] S. M. Marcora and W. Staiano. The limit to exercise tolerance in humans: mind over muscle? *European journal of applied physiology*, 109(4):763–770, 2010.
- [129] C. Martijn, P. Tenbült, H. Merckelbach, E. Dreezens, and N. K. de Vries. Getting a grip on ourselves: Challenging expectancies about loss of energy after self-control. *Social Cognition*, 20(6):441–460, 2002.
- [130] M. G. Mattar and N. D. Daw. Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience*, 21(11):1609, 2018.
- [131] J. L. McClelland, B. L. McNaughton, and R. 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.
- [132] C. McCormick, D. N. Barry, A. Jafarian, G. R. Barnes, and E. A. Maguire. vmPFC drives hippocampal processing during autobiographical memory recall regardless of remoteness. *bioRxiv*, 2020.

- [133] K. Mehlhorn, B. R. Newell, P. M. Todd, M. D. Lee, K. Morgan, V. A. Braithwaite, D. Hausmann, K. Fiedler, and C. Gonzalez. Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, 2(3):191, 2015.
- [134] C. Messier. Glucose improvement of memory: a review. *European journal of pharmacology*, 490(1-3):33–57, 2004.
- [135] J. Metcalfe and J. Xu. People mind wander more during massed than spaced inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(6):978, 2016.
- [136] D. E. Meyer and D. E. Kieras. A computational theory of executive cognitive processes and multiple-task performance: Part i. basic mechanisms. *Psychological review*, 104(1):3, 1997.
- [137] J. S. Mill. Utilitarianism. *Collected Works of John Stuart Mill*, 10, 1861.
- [138] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.
- [139] D. C. Molden, C. M. Hui, A. A. Scholer, B. P. Meier, E. E. Noreen, P. R. D’Agostino, and V. Martin. Motivational versus metabolic effects of carbohydrates on self-control. *Psychological science*, 23(10):1137–1144, 2012.
- [140] I. Momennejad, A. R. Otto, N. D. Daw, and K. A. Norman. Offline replay supports planning in human reinforcement learning. *Elife*, 7:e32548, 2018.
- [141] P. R. Montague, P. Dayan, and T. J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of neuroscience*, 16(5):1936–1947, 1996.
- [142] P. R. Montague, R. J. Dolan, K. J. Friston, and P. Dayan. Computational psychiatry. *Trends in cognitive sciences*, 16(1):72–80, 2012.
- [143] A. W. Moore and C. G. Atkeson. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine learning*, 13(1):103–130, 1993.
- [144] E. I. Moser, E. Kropff, and M.-B. Moser. Place cells, grid cells, and the brain’s spatial representation system. *Annu. Rev. Neurosci.*, 31:69–89, 2008.
- [145] T. Müller and M. A. Apps. Motivational fatigue: A neurocognitive framework for the impact of effortful exertion on subsequent motivation. *Neuropsychologia*, 123:141–151, 2019.

- [146] S. Musslick, B. Dey, K. Özçimder, M. M. A. Patwary, T. L. Willke, and J. D. Cohen. Controlled vs. automatic processing: A graph-theoretic approach to the analysis of serial vs. parallel processing in neural network architectures. In *CogSci*, 2016.
- [147] Z. Nádasdy, H. Hirase, A. Czurkó, J. Csicsvari, and G. Buzsáki. Replay and time compression of recurring spike sequences in the hippocampus. *Journal of Neuroscience*, 19(21):9497–9507, 1999.
- [148] J. Nakamura and M. Csikszentmihalyi. The concept of flow. In *Flow and the foundations of positive psychology*, pages 239–263. Springer, 2014.
- [149] D. Navon and D. Gopher. On the economy of the human-processing system. *Psychological review*, 86(3):214, 1979.
- [150] A. B. Newberg, J. Wang, H. Rao, R. L. Swanson, N. Wintering, J. S. Karp, A. Alavi, J. H. Greenberg, and J. A. Detre. Concurrent cbf and cmrglc changes during human brain activation by combined fmri–pet scanning. *Neuroimage*, 28(2):500–506, 2005.
- [151] W. Nicholson and C. M. Snyder. *Microeconomic theory: Basic principles and extensions*. Nelson Education, 2012.
- [152] R. K. Niyogi, Y.-A. Breton, R. B. Solomon, K. Conover, P. Shizgal, and P. Dayan. Optimal indolence: a normative microscopic approach to work and leisure. *Journal of The Royal Society Interface*, 11(91):20130969, 2014.
- [153] R. K. Niyogi, P. Shizgal, and P. Dayan. Some work and some play: microscopic and macroscopic approaches to labor and leisure. *PLoS computational biology*, 10(12):e1003894, 2014.
- [154] M. M. Nour, Y. Liu, C. Higgins, M. W. Woolrich, and R. J. Dolan. Reduced coupling between offline neural replay events and default mode network activation in schizophrenia. *Brain Communications*, 5(2):fcad056, 2023.
- [155] J. O’Keefe and J. Dostrovsky. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 1971.
- [156] H. F. Ólafsdóttir, D. Bush, and C. Barry. The role of hippocampal replay in memory and planning. *Current Biology*, 28(1):R37–R50, 2018.
- [157] A. R. Otto, S. J. Gershman, A. B. Markman, and N. D. Daw. The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological science*, 24(5):751–761, 2013.

- [158] A. R. Otto, C. M. Raio, A. Chiang, E. A. Phelps, and N. D. Daw. Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, 110(52):20941–20946, 2013.
- [159] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2):265–286, 2007.
- [160] P. Pajkossy, Á. Szöllösi, G. Demeter, and M. Racsomány. Tonic noradrenergic activity modulates explorative behavior and attentional set shifting: Evidence from pupillometry and gaze pattern analysis. *Psychophysiology*, 54(12):1839–1854, 2017.
- [161] N. Pattyn, X. Neyt, D. Henderickx, and E. Soetens. Psychophysiological investigation of vigilance decrement: boredom or cognitive fatigue? *Physiology & behavior*, 93(1-2):369–378, 2008.
- [162] J. Peng and R. J. Williams. Efficient learning and planning within the dyna framework. *Adaptive Behavior*, 1(4):437–454, 1993.
- [163] S. K. Perry. *Writing in flow: Keys to enhanced creativity*. Writers Digest Books, 1999.
- [164] J. C. Peterson, D. D. Bourgin, M. Agrawal, D. Reichman, and T. L. Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021.
- [165] B. E. Pfeiffer and D. J. Foster. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497(7447):74, 2013.
- [166] B. A. Poser, M. J. Versluis, J. M. Hoogduin, and D. G. Norris. Bold contrast sensitivity enhancement and artifact reduction with multiecho epi: parallel-acquired inhomogeneity-desensitized fmri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 55(6):1227–1235, 2006.
- [167] M. Posner and C. R. Snyder. Facilitation and inhibition. *Attention and performance*, 1975.
- [168] S. Posse, S. Wiese, D. Gembris, K. Mathiak, C. Kessler, M.-L. Grosse-Ruyken, B. Elghahwagi, T. Richards, S. R. Dager, and V. G. Kiselev. Enhancement of bold-contrast sensitivity by single-shot multi-echo functional mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(1):87–97, 1999.
- [169] M. E. Raichle and M. A. Mintun. Brain work and brain imaging. *Annu. Rev. Neurosci.*, 29:449–476, 2006.

- [170] D. Randles, I. Harlow, and M. Inzlicht. A pre-registered naturalistic observation of within domain mental fatigue and domain-general depletion of self-control. *PloS one*, 12(9):e0182980, 2017.
- [171] C. P. Rea and V. Modigliani. The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning: Journal of Practical Research & Applications*, 1985.
- [172] H. A. Ross, P. N. Russell, and W. S. Helton. Effects of breaks and goal switches on the vigilance decrement. *Experimental brain research*, 232(6):1729–1737, 2014.
- [173] E. M. Russek, I. Momennejad, M. M. Botvinick, S. J. Gershman, and N. D. Daw. Neural evidence for the successor representation in choice evaluation. *bioRxiv*, pages 2021–08, 2021.
- [174] D. D. Salvucci and N. A. Taatgen. Threaded cognition: An integrated theory of concurrent multitasking. *Psychological review*, 115(1):101, 2008.
- [175] R. Sayres and K. Grill-Spector. Relating retinotopic and object-selective responses in human lateral occipital cortex. *Journal of neurophysiology*, 100(1):249–267, 2008.
- [176] T. M. Scanlon. Contractualism and utilitarianism. 1982.
- [177] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.
- [178] A. C. Schapiro, E. A. McDevitt, T. T. Rogers, S. C. Mednick, and K. A. Norman. Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nature communications*, 9(1):3920, 2018.
- [179] A. C. Schapiro, N. B. Turk-Browne, M. M. Botvinick, and K. A. Norman. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160049, 2017.
- [180] T. Schaul, J. Quan, I. Antonoglou, and D. Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.
- [181] U. Schimmack. The ironic effect of significant results on the credibility of multiple-study articles. *Psychological methods*, 17(4):551, 2012.

- [182] J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.
- [183] W. Schneider and R. M. Shiffrin. Controlled and automatic human information processing: I. detection, search, and attention. *Psychological review*, 84(1):1, 1977.
- [184] N. W. Schuck and Y. Niv. Sequential replay of nonspatial task states in the human hippocampus. *Science*, 364(6447):eaaw5181, 2019.
- [185] W. Schultz. Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, 80(1):1–27, 1998.
- [186] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.
- [187] E. Schulz, R. Bhui, B. C. Love, B. Brier, M. T. Todd, and S. J. Gershman. Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences*, 116(28):13903–13908, 2019.
- [188] B. Seymour, N. Daw, P. Dayan, T. Singer, and R. Dolan. Differential encoding of losses and gains in the human striatum. *Journal of Neuroscience*, 27(18):4826–4831, 2007.
- [189] C. E. Sezener, A. Dezfouli, and M. Keramati. Optimizing the depth and the direction of prospective planning using information values. *PLoS computational biology*, 15(3):e1006827, 2019.
- [190] A. Shenhav, M. M. Botvinick, and J. D. Cohen. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, 79(2):217–240, 2013.
- [191] R. M. Shiffrin and W. Schneider. Controlled and automatic human information processing: Ii. perceptual learning, automatic attending and a general theory. *Psychological review*, 84(2):127, 1977.
- [192] J. D. Shin, W. Tang, and S. P. Jadhav. Dynamics of awake hippocampal-prefrontal replay for spatial learning and memory-guided decision making. *Neuron*, 2019.
- [193] H. A. Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118, 1955.
- [194] C. A. Sims. Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690, 2003.

- [195] J. Smallwood and J. W. Schooler. The restless mind. *Psychological bulletin*, 132(6):946, 2006.
- [196] J. Smallwood and J. W. Schooler. The science of mind wandering: empirically navigating the stream of consciousness. *Annual review of psychology*, 66:487–518, 2015.
- [197] A. Solway and M. M. Botvinick. Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychological review*, 119(1):120, 2012.
- [198] R. S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4):160–163, 1991.
- [199] R. S. Sutton, A. G. Barto, et al. *Introduction to reinforcement learning*, volume 2. MIT press Cambridge, 1998.
- [200] A. Tambini, N. Ketz, and L. Davachi. Enhanced brain correlations during rest are related to memory for recent experiences. *Neuron*, 65(2):280–290, 2010.
- [201] E. Thorndike. Mental fatigue. i. *Psychological Review*, 7(6):547, 1900.
- [202] E. L. Thorndike. The curve of work. *Psychological Review*, 19(3):165, 1912.
- [203] G. Tononi and C. Cirelli. Sleep and synaptic homeostasis: a hypothesis. *Brain research bulletin*, 62(2):143–150, 2003.
- [204] G. Tononi and C. Cirelli. Sleep function and synaptic homeostasis. *Sleep medicine reviews*, 10(1):49–62, 2006.
- [205] G. Tononi and C. Cirelli. Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration. *Neuron*, 81(1):12–34, 2014.
- [206] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.
- [207] G. M. van de Ven, S. Trouche, C. G. McNamara, K. Allen, and D. Dupret. Hippocampal offline reactivation consolidates recently formed cell assembly patterns during sharp wave-ripples. *Neuron*, 92(5):968–974, 2016.
- [208] D. Van der Linden, M. Frese, and T. F. Meijman. Mental fatigue and the control of cognitive processes: effects on perseveration and planning. *Acta psychologica*, 113(1):45–65, 2003.
- [209] M. A. Van Der Meer and A. D. Redish. Covert expectation-of-reward in rat ventral striatum at decision points. *Frontiers in integrative neuroscience*, 3:1, 2009.

- [210] E. J. Wamsley. Memory consolidation during waking rest. *Trends in cognitive sciences*, 23(3):171–173, 2019.
- [211] X.-J. Wang and J. H. Krystal. Computational psychiatry. *Neuron*, 84(3):638–654, 2014.
- [212] J. S. Warm, R. Parasuraman, and G. Matthews. Vigilance requires hard mental work and is stressful. *Human factors*, 50(3):433–441, 2008.
- [213] C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.
- [214] A. M. Wikenheiser and A. D. Redish. Hippocampal theta sequences reflect current goals. *Nature neuroscience*, 18(2):289, 2015.
- [215] R. Wild and A. Hill. *Women in the factory: A study of job satisfaction and labour turnover*. Institute of Personnel Management, 1970.
- [216] K. J. Williams, K. E. Lee, T. Hartig, L. D. Sargent, N. S. Williams, and K. A. Johnson. Conceptualising creativity benefits of nature experience: Attention restoration and mind wandering as complementary processes. *Journal of Environmental Psychology*, 59:36–45, 2018.
- [217] M. A. Wilson and B. L. McNaughton. Reactivation of hippocampal ensemble memories during sleep. *Science*, 265(5172):676–679, 1994.
- [218] R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, and J. D. Cohen. Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143(6):2074, 2014.
- [219] R. C. Wilson, A. Shenhav, M. Straccia, and J. D. Cohen. The eighty five percent rule for optimal learning. *Nature communications*, 10(1):1–9, 2019.
- [220] G. E. Wimmer, Y. Liu, N. Vehar, T. E. Behrens, and R. J. Dolan. Episodic memory retrieval is supported by rapid replay of episode content. *bioRxiv*, page 758185, 2019.
- [221] B. C. Wittmann, N. D. Daw, B. Seymour, and R. J. Dolan. Striatal activity underlies novelty-based choice in humans. *Neuron*, 58(6):967–973, 2008.
- [222] R. M. Yerkes and J. D. Dodson. The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology and psychology*, 18(5):459–482, 1908.
- [223] J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, and J. R. Reynolds. Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273, 2007.

- [224] C. M. Zedelius and J. W. Schooler. Mind wandering “ahas” versus mindful reasoning: alternative routes to creative solutions. *Frontiers in Psychology*, 6:834, 2015.

Appendix A

Appendix

A.1 Theory Methods

A.1.1 Cognitive Fatigue

$$EVB_C(s, s_k, a_k) = \mathbb{E}_{\pi_{\text{new}}} \left[\sum_{i=\tau}^{\infty} \gamma^i R_{t+i} \middle| S_t = s \right] - \mathbb{E}_{\pi_{\text{old}}} \left[\sum_{i=0}^{\infty} \gamma^i R_{t+i} \middle| S_t = s \right] \quad (\text{A.1})$$

$$= \gamma^\tau v_{\pi_{\text{new}}}(s) - v_{\pi_{\text{old}}}(s) \quad (\text{A.2})$$

$$= \gamma^\tau v_{\pi_{\text{new}}}(s) - \gamma^\tau v_{\pi_{\text{old}}}(s) + \gamma^\tau v_{\pi_{\text{old}}}(s) - v_{\pi_{\text{old}}}(s) \quad (\text{A.3})$$

$$= \gamma^\tau (v_{\pi_{\text{new}}}(s) - v_{\pi_{\text{old}}}(s)) + \gamma^\tau v_{\pi_{\text{old}}}(s) - v_{\pi_{\text{old}}}(s) \quad (\text{A.4})$$

$$= \gamma^\tau EVB(s, s_k, a_k) + \gamma^\tau v_{\pi_{\text{old}}}(s) - v_{\pi_{\text{old}}}(s) \quad (\text{A.5})$$

$$= \gamma^\tau EVB(s, s_k, a_k) - (1 - \gamma^\tau) v_{\pi_{\text{old}}}(s) \quad (\text{A.6})$$

$$= \gamma^\tau EVB(s, s_k, a_k) - (1 - \gamma^\tau) \sum_a \pi_{\text{old}}(a|s) q_{\pi_{\text{old}}}(s, a) \quad (\text{A.7})$$

A previous version of the manuscript incorrectly decomposed EVB_C into $Gain \times Need$ in which

$$Gain(s_k, a_k) = \gamma^\tau \sum_{a \in A} Q_{\pi_{\text{new}}}(s_k, a) \pi_{\text{new}}(a|s_k) - \sum_{a \in A} Q_{\pi_{\text{new}}}(s_k, a) \pi_{\text{old}}(a|s_k) \quad (\text{A.8})$$

This mistake has been corrected in the present manuscript.

The original gridworld comprised of a 9x6 maze, in which the agent started at (0,3) and there was a reward of 1 at (8,5). Walls were additionally located at (2,2), (2,3), (2,4), (7,3), (7,4), (7,5), and (5,1). The agent's Q-values and reward prior were both initialized to 0.

An agent’s policy is calculated using the softmax choice rule:

$$\pi(a|s) = \frac{e^{\beta Q(s,a)}}{\sum_{a'} e^{\beta Q(s,a')}} \quad (\text{A.9})$$

in which β is the inverse temperature. β was set to 5 for all simulations, unless otherwise stated.

After every state-action-reward-state transition, the agent updates its Q-values according to the temporal difference learning rule in Eq. 2.1. The learning rate α , was set to 0.90 for all simulations. Each replayed backup follow the same update equation. After every actually realized state-action-state transition (i.e., not replayed transitions), the successor representation also updated according the temporal difference learning rule.

At each time step, the $EV B_C$ is computed. If the value is greater than zero, the agent replays $\arg \max EV B_C$. Once it falls below zero, the agent samples from its policy π .

The easier gridworld comprised of a 4x6 maze, in which the agent started at (0,3) and there was a reward at (3,5). Walls were additionally located at (2,2), (2,3), and (2,4). All other details were the same as the agent in the harder gridworld.

The agent completed twenty episodes for the harder gridworld and forty episodes for the easier gridworld. Ten different runs were completed for each value of τ , and number of replay events and average reward rate were averaged over these runs.

The bottom panels in Figures 2.2 and Figures 2.3 plot the replay behavior of the ten different agent runs when $\tau = 0.04$. The right plot was a smoothed version of the left plot, using a Gaussian filter with $\sigma = 5$.

A.1.2 Boredom

The right panel of Figure 2.5 demonstrates how the Q_{VOI} of a dynamic task changes over time. An infinite horizon and a discount factor of $\gamma = 0.9$ were used for simulations. The q_{new} was calculated using the temporal difference learning update rule.

Figure 2.6 contrasts the exploratory behavior of an agent in a short horizon versus that in a long horizon. The uncertain action had a relative mean value represented by the x-axis, but was given uncertainty by simulating a fifty percent chance of changing by plus one and fifty percent chance of changing by minus one for twenty iterations. For the short horizon, a horizon of two was used, while a horizon of ten was chosen for the long condition. A discount factor of 0.5 was used for this simulation.

Furthermore, in order to keep the scaling of the Q-values consistent, the policy was computed with respect to the expected one-step reward (which was computed by dividing the Q-value by the Need term). Because of the nature of the bandit task, the Need term for both simulations were calculated analytically instead of

using a successor representation.

A.1.3 Replaying to Explore

The T-Maze was constructed in the Gridworld environment, using a 5x5 grid with walls everywhere except for the middle row and the last column. The agent started at (0,2), and there were two rewards, 1 and 5, at locations (4,0) and (4,4).

Rewards were shuffled randomly every five trials. A successor representation based on a uniform policy π was used instead of a dynamically updated successor representation. The agent's Q-values were initialized to 0, but the VOI's were initialized to $U_0 = 5$. For the Need term in Eq. 2.22, three different values of k were used: 0, 0.05, and 0.5. A value of $\tau = 0.04$ was used for all simulations. Forty simulations of forty episodes were run for each value of k .

Table A.1: Four-Way Classification of Category (incl. Rest) Within Runs

ROIs	Train Mean	Train SEM	Test Mean	Test SEM
22	0.656	0.012	0.513	0.009
23, 24	0.649	0.013	0.541	0.010
22, 23, 24	0.793	0.012	0.581	0.010
35, 36	0.499	0.011	0.441	0.008
22, 35, 36	0.756	0.012	0.540	0.009
38, 39, 40, 41	0.631	0.016	0.516	0.011
22, 35, 36, 38, 39, 40, 41	0.834	0.012	0.561	0.011
22, 23, 24, 35, 36, 38, 39, 40, 41	0.870	0.012	0.567	0.012

Table A.2: Four-Way Classification of Category (incl. Rest) Across Runs

ROIs	Train Mean	Train SEM	Test Mean	Test SEM
22	0.626	0.014	0.460	0.009
23, 24	0.663	0.015	0.532	0.012
22, 23, 24	0.773	0.012	0.564	0.010
35, 36	0.511	0.016	0.416	0.010
22, 35, 36	0.738	0.012	0.518	0.009
38, 39, 40, 41	0.613	0.017	0.481	0.013
22, 35, 36, 38, 39, 40, 41	0.808	0.012	0.573	0.010
22, 23, 24, 35, 36, 38, 39, 40, 41	0.844	0.012	0.605	0.010

A.2 Aggregate Classifier Results

Tables A.1 to A.6 report classifier results averaged across all subjects. For each subject, a 3 to 5 TR offset was used to map between input and output, with the selected offset being the one that optimized predictive accuracy across all ROI's.

Table A.3: Three-Way Classification of Category Within Runs

ROIs	Train Mean	Train SEM	Test Mean	Test SEM
22	0.661	0.010	0.468	0.011
23, 24	0.585	0.014	0.436	0.010
22, 23, 24	0.781	0.010	0.503	0.011
35, 36	0.415	0.008	0.326	0.005
22, 35, 36	0.739	0.010	0.456	0.011
38, 39, 40, 41	0.569	0.015	0.406	0.011
22, 35, 36, 38, 39, 40, 41	0.830	0.009	0.478	0.011
22, 23, 24, 35, 36, 38, 39, 40, 41	0.875	0.010	0.481	0.012

Table A.4: Three-Way Classification of Category Across Runs

ROIs	Train Mean	Train SEM	Test Mean	Test SEM
22	0.674	0.008	0.427	0.009
23, 24	0.670	0.008	0.462	0.009
22, 23, 24	0.764	0.009	0.480	0.010
35, 36	0.628	0.007	0.377	0.005
22, 35, 36	0.731	0.009	0.438	0.009
38, 39, 40, 41	0.678	0.008	0.466	0.008
22, 35, 36, 38, 39, 40, 41	0.805	0.009	0.488	0.010
22, 23, 24, 35, 36, 38, 39, 40, 41	0.846	0.009	0.524	0.010

Table A.5: Binary Classification of Category Within Runs

ROIs	Category	Train Mean	Train SEM	Test Mean	Test SEM
22	Face	0.834	0.012	0.751	0.011
22	Place	0.859	0.010	0.780	0.010
22	Object	0.867	0.009	0.800	0.009
23, 24	Face	0.892	0.012	0.851	0.013
23, 24	Place	0.887	0.011	0.847	0.013
23, 24	Object	0.888	0.010	0.849	0.010
22, 23, 24	Face	0.927	0.009	0.849	0.012
22, 23, 24	Place	0.925	0.009	0.850	0.011
22, 23, 24	Object	0.923	0.008	0.848	0.009
35, 36	Face	0.814	0.016	0.769	0.014
35, 36	Place	0.829	0.016	0.783	0.015
35, 36	Object	0.808	0.016	0.762	0.015
22, 35, 36	Face	0.913	0.009	0.817	0.011
22, 35, 36	Place	0.918	0.009	0.828	0.011
22, 35, 36	Object	0.913	0.008	0.829	0.009
38, 39, 40, 41	Face	0.882	0.012	0.828	0.013
38, 39, 40, 41	Place	0.874	0.013	0.818	0.014
38, 39, 40, 41	Object	0.869	0.013	0.815	0.013
22, 35, 36, 38, 39, 40, 41	Face	0.941	0.009	0.840	0.011
22, 35, 36, 38, 39, 40, 41	Place	0.937	0.009	0.839	0.012
22, 35, 36, 38, 39, 40, 41	Object	0.934	0.008	0.836	0.010
22, 23, 24, 35, 36, 38, 39, 40, 41	Face	0.950	0.008	0.848	0.012
22, 23, 24, 35, 36, 38, 39, 40, 41	Place	0.945	0.008	0.845	0.012
22, 23, 24, 35, 36, 38, 39, 40, 41	Object	0.943	0.008	0.842	0.010

Table A.6: Binary Classification of Category Across Runs

ROIs	Category	Train Mean	Train SEM	Test Mean	Test SEM
22	Face	0.802	0.013	0.706	0.012
22	Place	0.827	0.012	0.727	0.012
22	Object	0.846	0.010	0.763	0.012
23, 24	Face	0.865	0.014	0.820	0.015
23, 24	Place	0.865	0.014	0.823	0.014
23, 24	Object	0.863	0.012	0.821	0.014
22, 23, 24	Face	0.916	0.010	0.844	0.011
22, 23, 24	Place	0.918	0.009	0.843	0.011
22, 23, 24	Object	0.917	0.009	0.844	0.010
35, 36	Face	0.750	0.015	0.702	0.013
35, 36	Place	0.768	0.016	0.715	0.014
35, 36	Object	0.745	0.015	0.699	0.013
22, 35, 36	Face	0.897	0.010	0.794	0.010
22, 35, 36	Place	0.904	0.010	0.814	0.011
22, 35, 36	Object	0.904	0.009	0.817	0.010
38, 39, 40, 41	Face	0.828	0.014	0.771	0.014
38, 39, 40, 41	Place	0.823	0.014	0.754	0.015
38, 39, 40, 41	Object	0.815	0.014	0.749	0.014
22, 35, 36, 38, 39, 40, 41	Face	0.929	0.009	0.851	0.010
22, 35, 36, 38, 39, 40, 41	Place	0.927	0.009	0.852	0.011
22, 35, 36, 38, 39, 40, 41	Object	0.928	0.008	0.845	0.010
22, 23, 24, 35, 36, 38, 39, 40, 41	Face	0.940	0.009	0.869	0.010
22, 23, 24, 35, 36, 38, 39, 40, 41	Place	0.937	0.009	0.867	0.011
22, 23, 24, 35, 36, 38, 39, 40, 41	Object	0.937	0.008	0.860	0.010