The Temporal Dynamics of Opportunity Costs: A Normative Account of Cognitive Fatigue and Boredom

Mayank Agrawal*
Princeton Neuroscience Institute and Department of Psychology, Princeton University
*Corresponding Author: mayank.agrawal@princeton.edu

Marcelo G. Mattar
Department of Cognitive Science, UC San Diego

Jonathan D. Cohen†
Princeton Neuroscience Institute and Department of Psychology, Princeton University

Nathaniel D. Daw†
Princeton Neuroscience Institute and Department of Psychology, Princeton University

†: The two co-last authors, listed alphabetically, made equal and complementary contributions to the guidance and support of the work reported in this article.

Abstract

Cognitive fatigue and boredom are two phenomenological states widely associated with limitations in cognitive control. In this paper, we present a rational analysis of the temporal structure of controlled behavior, which provides a new framework for providing 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, rather then 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 exploratory actions (external in the world) to gather information. 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.

Keywords: cognitive control, cognitive fatigue, boredom, hippocampal replay, explore-exploit, reinforcement learning, opportunity costs

The Temporal Dynamics of Opportunity Costs: A Normative Account of Cognitive Fatigue and Boredom

Contents

Abstract 2

The Temporal Dynamics of Opportunity Costs: A Normative Account of Cognitive Fatigue and Boredom 3

Introduction **6**

Overview 7

Roadmap **8**

Background **9**

Rational Models of Cognitive Control 9

Reinforcement Learning 10

Model-Free vs. Model-Based Learning 11

Exploration vs Exploitation 12

Part 1: Cognitive Fatigue **13**

Empirical Findings 14

1. Rest Helps Performance 14

2. Arbitration Between Labor and Leisure 15

3. Difficult Tasks are More Fatiguing 15

Discussion 16

Hippocampal Replay 16

Mattar and Daw (2018) 17

Expected Value of Backup with Cost (EVB_C). 18

Hippocampal Replay as the Value of Leisure 19

Results . . . 21
Discussion . . . 23
Benefits and Limitations of Replay . . . 24
Intra-Trial Dynamics . . . 24
A Normative Lens to Understand Psychiatric Illnesses . . . 25
Hippocampal Lesions . . . 26

Part 2: Boredom **27**
Arousal and Boredom . . . 28
Empirical Findings . . . 29
1. Optimal Participant-Task Fit . . . 29
2. Switching to Other Tasks . . . 30
Discussion . . . 31
Formalizing the Value of Information . . . 32
Discussion . . . 35
Adaptive Gain Theory . . . 36
Curiosity . . . 36

Part 3: Replaying to Explore **37**
Simulations . . . 38
Model . . . 38
Results . . . 39
Discussion . . . 40

General Discussion **41**
Limitations & Future Directions . . . 43
Algorithmic Approximations . . . 43
Causal Disruptions of Fatigue . . . 43
Conclusion . . . 44

Acknowledgements **44**

References **45**

Appendix **60**

Methods 60

Part 1: Cognitive Fatigue 60

Part 2: Boredom 61

Part 3: Replaying to Explore 61

Introduction

Learning 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 (Posner & Snyder, 1975; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977; Anderson, 1987). 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 (Sutton, Barto, et al., 1998; J. D. Cohen, Dunbar, & McClelland, 1990; Daw, Niv, & Dayan, 2005).

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 accord 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 account of 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 weeks[1]. This is unlikely to be true (Arai, 1912; Huxtable, White, & McCartor, 1946; N. H. Mackworth, 1948; Van der Linden, Frese, & Meijman, 2003; Healy, Kole, Buck-Gengler, & Bourne, 2004; Lorist, Boksem, & Ridderinkhof, 2005; Warm, Parasuraman, & Matthews, 2008). 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

[1] While the effects of massing vs. spacing of practice have been studied in several contexts, and several relevant factors have been identified (Izawa, 1971; Bloom & Shuell, 1981; Rea & Modigliani, 1985; Donovan & Radosevich, 1999; Metcalfe & Xu, 2016), fatigue and/or boredom are almost certain to be dominant ones in this example.

resource (e.g., metabolic; Baumeister, Bratslavsky, Muraven, & Tice, 1998; Baumeister & Vohs, 2007), by analogy to the case of physical fatigue. Boredom, in turn, has been attributed to a diminution of motivation. In contrast to these views, Kurzban, Duckworth, Kable, and Myers (2013) proposed that exerting cognitive control (limitations of which may potentially characterize fatigue and boredom) may instead be accompanied with 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. However, this broad approach admits of 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.

Overview

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. 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.

Roadmap

Part 1 proposes 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. Part 2 proposes that boredom tracks the value of exploring alternate tasks, here playing out via competition between different classes of overt action: information-seeking vs. exploitative. We offer a model of directed exploration for navigating 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, Part 3 integrates 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 1.

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

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

Figure 1. This paper partitions the constraints associated with the duration of cognitive control into two dimensions: action vs. replay and exploration vs. exploitation. Part 1 investigates fatigue, which we propose to be a signal for the value replay while Part 2 investigates boredom, which we propose to be a bias towards exploration. Part 3 integrates the mechanisms of replay and exploration; whether this corresponds to a currently studied phenomenology is unknown and is an important direction for future research.

Background

Rational Models of Cognitive Control

Cognitive fatigue has long been studied in psychology (E. Thorndike, 1900; Dodge, 1917), with one influential account, 'ego depletion,' suggesting that it reflects the consumption and subsequent diminution of a metabolic resource (Baumeister et al., 1998; Baumeister & Vohs, 2007). Recently, however, this hypothesis has been called into question (Kurzban et al., 2013); and multiple meta-analyses have challenged its empirical foundation, suggesting that the associated studies overestimated null effects (Carter, Kofler, Forster, & McCullough, 2015;

Hagger et al., 2016; Randles, Harlow, & Inzlicht, 2017). 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 (Messier, 2004; Raichle & Mintun, 2006; Gibson, 2007; Molden et al., 2012; Schimmack, 2012). In fact, some of the largest glucose demands in the brain arise from visual processing (Newberg et al., 2005), 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 (Shenhav, Botvinick, & Cohen, 2013; Kurzban et al., 2013; Lieder & Griffiths, 2020) propose that performance variation arises from rational balancing of the costs vs. 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.

The current theory instantiates a case of the Expected Value of Control (EVC) framework (Shenhav et al., 2013), as applied to fatigue and boredom. Agents' actions in our model span two dimensions: physical vs. mental and exploratory vs. exploitative. 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.

Reinforcement Learning

Reinforcement learning (Sutton et al., 1998; Daw et al., 2005) offers an integrative computational framework in which to implement an EVC model. 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.

Model-Free vs. Model-Based Learning. Two main classes of algorithms have emerged in reinforcement learning: model-free and model-based learning (Sutton et al., 1998; Daw et al., 2005). 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 (Watkins & Dayan, 1992) estimate this function directly from experienced rewards over experienced state trajectories (Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997; Schultz, 1998). 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:

$$Q(s, a) \leftarrow Q(s, a) + \alpha(r + \gamma \max_{a'} Q(s', a') - Q(s, a)) \tag{1}$$

in which α refers to the learning rate of the agent. In contrast, model-based learning (Daw et al., 2005; Solway & Botvinick, 2012) 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. 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). For these reasons this computational distinction has been employed in neuroscience as a model of the tradeoffs between thinking and acting (Daw et al., 2005; Keramati, Dezfouli, & Piray, 2011). 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 (Daw et al., 2005; Keramati et al., 2011; Otto, Gershman, Markman, & Daw, 2013; Otto, Raio, Chiang, Phelps, & Daw, 2013; S. W. Lee, Shimojo, & O'Doherty, 2014; Keramati, Smittenaar, Dolan, & Dayan, 2016; Kool, Cushman, & Gershman, 2016; Kool, Gershman, & Cushman, 2017; Sezener, Dezfouli, & Keramati, 2019).

The current theory employs a finer grained version of this idea, based on the Dyna framework (Sutton, 1991), which learns values from experience using Eq. 1, but also makes decisions about whether to improve these using individual steps of model evaluation and, if so, which ones (Mattar & Daw, 2018). 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 and Daw (2018) proposed that the brain implements the steps of model evaluation by replaying trajectories in hippocampus. This theory will be the basis of our analysis in Part 1, where we suggest fatigue tracks the value of such replay.

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 (Kaelbling, Littman, & Moore, 1996; Sutton et al., 1998; Daw, O'doherty, Dayan, Seymour, & Dolan, 2006; J. D. Cohen, McClure, & Yu, 2007; R. C. Wilson, Geana, White, Ludvig, & Cohen, 2014; Mehlhorn et al., 2015; Gershman, 2018; Schulz et al., 2019).

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

(Gittins, 1979) 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, vs. 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):

$$Q_{VOI}(s, a) = Q(s, a) + VOI(s, a) \tag{2}$$

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 (Auer, Cesa-Bianchi, & Fischer, 2002), which proposes the VOI in a multi-armed bandit setting to be

$$VOI = \sqrt{\frac{2 \ln n}{n_i}} \tag{3}$$

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 Part 2, where we suggest boredom reflects a bias towards exploration.

Part 1: 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 (Kool & Botvinick, 2014), with Edward Thorndike defining fatigue as "that diminution in efficiency *which rest can cure* (emphasis ours)" over one hundred years ago (E. L. Thorndike, 1912). 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.

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.

1. Rest Helps Performance. Several studies have examined the effects of rest in mitigating decrements in performance with time on task (Bergum & Lehr, 1962; Ross, Russell, & Helton, 2014; Helton & Russell, 2015, 2017). For example, Helton and Russell (2015) demonstrated the benefit of rest by having participants carry out a vigilance task (N. H. Mackworth, 1948) 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, Helton and Russell (2015) 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 was 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.

2. Arbitration Between Labor and Leisure. Kool and Botvinick (2014) 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 (Nicholson & Snyder, 2012). 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 (Kirchner, 1958) 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 (Shenhav et al., 2013) However, this account did not provide an explanation for the value of rest.

3. 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, Blain, Hollard, and Pessiglione (2016) conducted a study over the span of six hours. Participants performed either the ‘easy’ tasks of a one-back and one-switch[2] 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

[2] In an N-switch block, the participant switches between two tasks N times.

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.

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 (McClelland, McNaughton, & O'Reilly, 1995; Tambini, Ketz, & Davachi, 2010; Carr, Jadhav, & Frank, 2011; Ólafsdóttir, Bush, & Barry, 2018; Wamsley, 2019). 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 action[3]. 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.

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 (O'Keefe & Dostrovsky, 1971; Moser, Kropff, & Moser, 2008). 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 (Foster & Wilson, 2006; Diba & Buzsáki, 2007; Davidson, Kloosterman, & Wilson, 2009; Karlsson & Frank, 2009; Gupta, van der Meer, Touretzky, & Redish, 2010). Replay events are commonly observed during epochs of high-frequency

[3] A similar argument has been made for sleep, suggesting that sleep is the 'price that the brain pays for plasticity' (Tononi & Cirelli, 2003, 2006, 2014)

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 (M. A. Wilson & McNaughton, 1994; Nádasdy, Hirase, Czurkó, Csicsvari, & Buzsáki, 1999; Louie & Wilson, 2001; A. K. Lee & Wilson, 2002). 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. (Gershman, Markman, & Otto, 2014; Schapiro, McDevitt, Rogers, Mednick, & Norman, 2018; Momennejad, Otto, Daw, & Norman, 2018; Liu, Dolan, Kurth-Nelson, & Behrens, 2019; Wimmer, Liu, Vehar, Behrens, & Dolan, 2019; Schuck & Niv, 2019; Eldar, Lièvre, Dayan, & Dolan, 2020).

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.

Mattar and Daw (2018). Mattar and Daw (2018) (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 (Sutton et al., 1998; Daw et al., 2005). 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 (Watkins & Dayan, 1992), 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 (Sutton, 1991). If this simulated experience is faster than physical experience and/or selected through a priority metric (Peng & Williams, 1993; Moore & Atkeson, 1993), the agent can converge to the optimal policy quicker, and thus increase future reward, as opposed to the 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 with the highest EVB at the beginning and end of a trial. They 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 (Peng & Williams, 1993; Moore & Atkeson, 1993; Schaul, Quan, Antonoglou, & Silver, 2015), 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, $EVB_C(s_k, a_k)$ is the expected increase in reward resulting from replaying the state-action pair (s_k, a_k) and executing the corresponding Bellman backup (i.e. temporal difference learning update, as in Equation 1 but for a simulated rather than experienced step), discounted by the amount of reward lost due to the time it takes to replay. As in the M&D model, $EVB_C(s_k, a_k)$ can be decomposed into the product of *Gain* and *Need* terms.

$$EVB_C(s_k, a_k) = \mathop{\mathbb{E}}_{\pi_{\text{new}}}\left[\sum_{i=\tau}^{\infty}\gamma^i R_{t+1}\middle| S_t = s\right] - \mathop{\mathbb{E}}_{\pi_{\text{old}}}\left[\sum_{i=0}^{\infty}\gamma^i R_{t+1}\middle| S_t = s\right] \tag{4}$$

$$= Gain(s_k, a_k) \times Need(s_k) \tag{5}$$

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). The *Need* term corresponds to the expected number of (delay discounted) future visits to that state, which corresponds to the successor representation (Dayan, 1993; Gershman, Moore, Todd, Norman, & Sederberg, 2012). Here, we extend the M&D model by including a cost term, scaled by τ which is the ratio of time it takes to replay versus act[4].

Thus, in the current variant, the components of EVB_C can be expressed as:

$$Gain(s_k, a_k) = \gamma^{\tau}\sum_{a\in A} Q_{\pi_{\text{new}}}(s_k, a)\pi_{\text{new}}(a|s_k) - \sum_{a\in A} Q_{\pi_{\text{new}}}(s_k, a)\pi_{\text{old}}(a|s_k) \tag{6}$$

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 (and discounted for a delay of τ), vs. following the status quo policy π_{old}; and

$$Need(s_k) = \mathbb{E}\left[\sum_{i=0}^{\infty}\gamma^i \delta_{s_{t+i}, s_k} \mid s_t\right] = M(s_t, s_k) \tag{7}$$

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_t. This in turn can be obtained from the successor representation M, estimated as $\widehat{M}$, for the state pair (s_t, s_k).

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

[4] One 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 (Pfeiffer & Foster, 2013) as compared to the speed of running on a track, which is approximately 40 cm/s (Wikenheiser & Redish, 2015). In humans, a recent study by Wimmer et al. (2019) suggests that replay can be sixty times faster than physical action.

these rest states. Defining EVB_C^* as $\max EVB_C$, a rational agent should replay the most valuable location, $\arg\max EVB_C$, as long as its value $EVB_C^* > 0$. Hence, the overall value of rest — which we propose is subjectively sensed as fatigue — is EVB_C^*. Once this value falls below zero, the agent should physically act. 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 (Kool & Botvinick, 2014; Niyogi, Breton, et al., 2014; Niyogi, Shizgal, & Dayan, 2014; Inzlicht, Schmeichel, & Macrae, 2014; Dora, van Hooff, Geurts, Kompier, & Bijleveld, 2019).

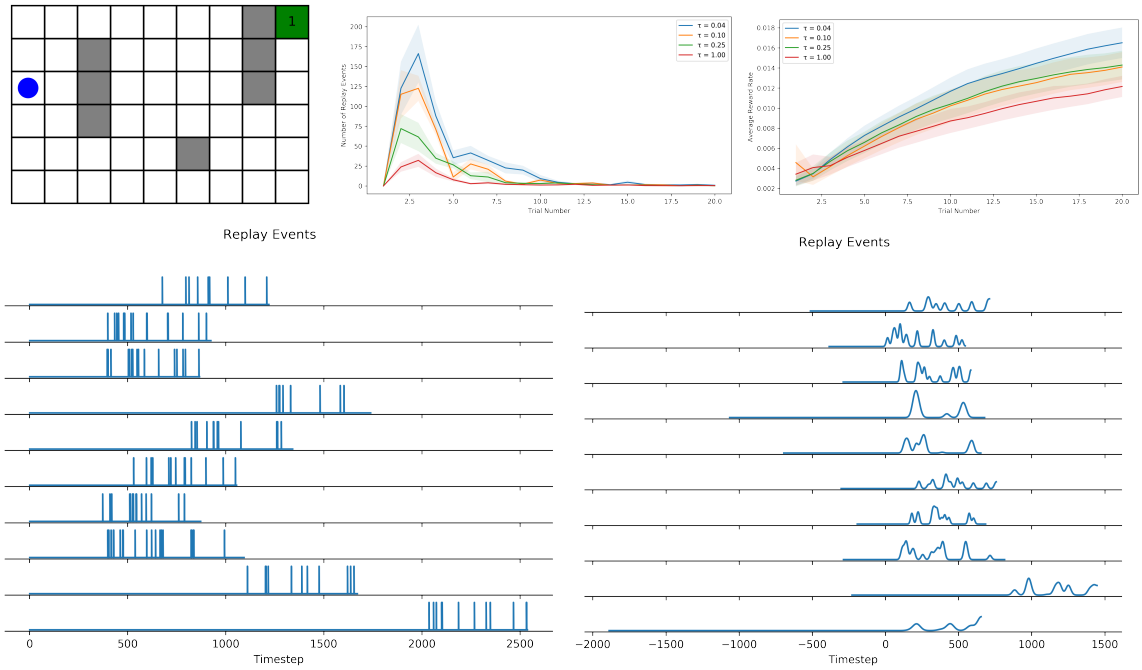

Figure 2. Simulation results for gridworld agent. (Top Left) Gridworld environment used for simulations. (Top Middle) Number of replay events over the course of multiple trials for different values of τ. (Top Right) Average reward rate for different values of τ. All error bars indicate ± 1 SEM. (Bottom 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).

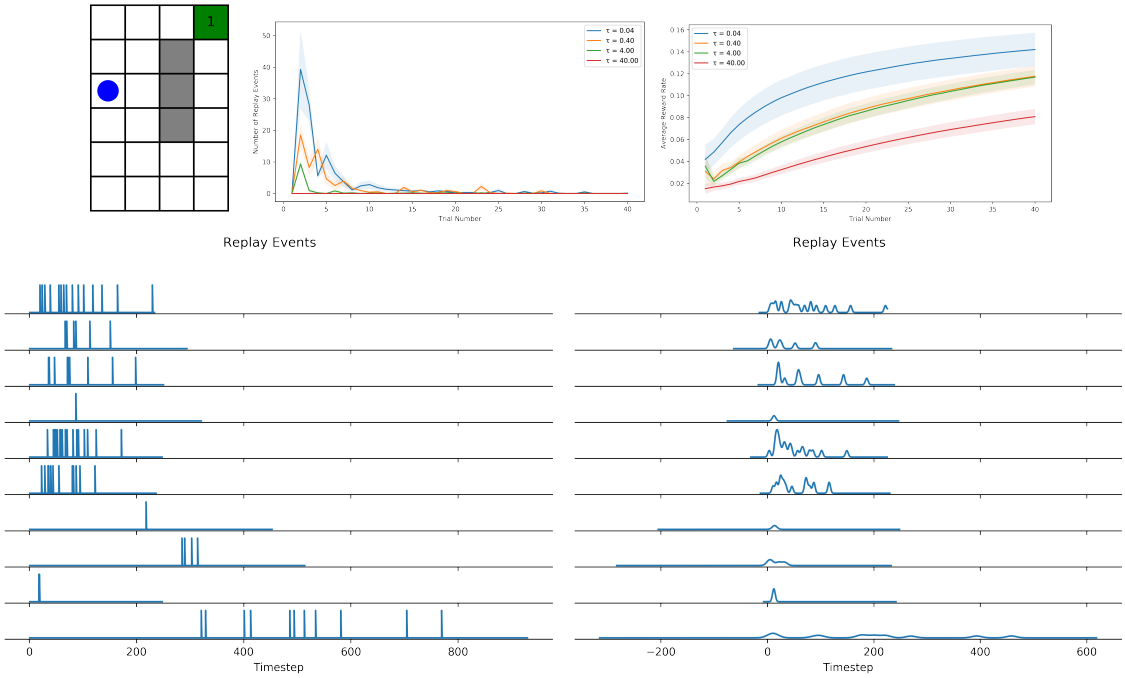

Figure 3. Simulation results for running easy gridworld agent. (Top Left) Easy gridworld environment used for simulations. (Top Middle) Number of replay events over the course of multiple trials for different values of τ. (Top Right) Average reward rate for different values of τ. All error bars indicate ± 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).

Results. Figure 2 plots the replay behavior of an agent pursuing a specified reward in a gridworld, using the EVB_C-driven replay algorithm descried 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 it 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; in this regime, gain (and EVB_C) are positive. 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 opportunity costs of replay exceed its benefits (Van Der Meer & Redish, 2009; van de Ven, Trouche, McNamara, Allen, & Dupret, 2016). This third regime — in which the value function has converged, gain 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 Helton and Russell (2015) 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 top right panel in Figure 2 demonstrates that agents that replay more perform better. 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 (Navon & Gopher, 1979; Meyer & Kieras, 1997; Salvucci & Taatgen, 2008; Musslick et al., 2016). 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 also grounds the labor-leisure tradeoff of Kool and Botvinick (2014) in normative terms, if it is assumed that leisure corresponds to time allocated for replay. Rather than posit leisure as intrinsically valuable, we demonstrate how the temporal dynamics of its value as an

opportunity for relay (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 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 Blain et al. (2016). 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 Blain et al. (2016).

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 Helton and Russell (2015) 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 (McClelland et al., 1995; Kumaran, Hassabis, & McClelland, 2016; Schapiro, Turk-Browne, Botvinick, & Norman, 2017) 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 (Mnih et al., 2015). 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. Dasgupta, Smith, Schulz, Tenenbaum, and Gershman (2018) 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 (Carr et al., 2011; Ólafsdóttir et al., 2018).The dynamics of the EVB_C agent are shown in Figure 4, and thus we predict that sequential tasks should have specific patterns of fatigue dynamics within individual trials.

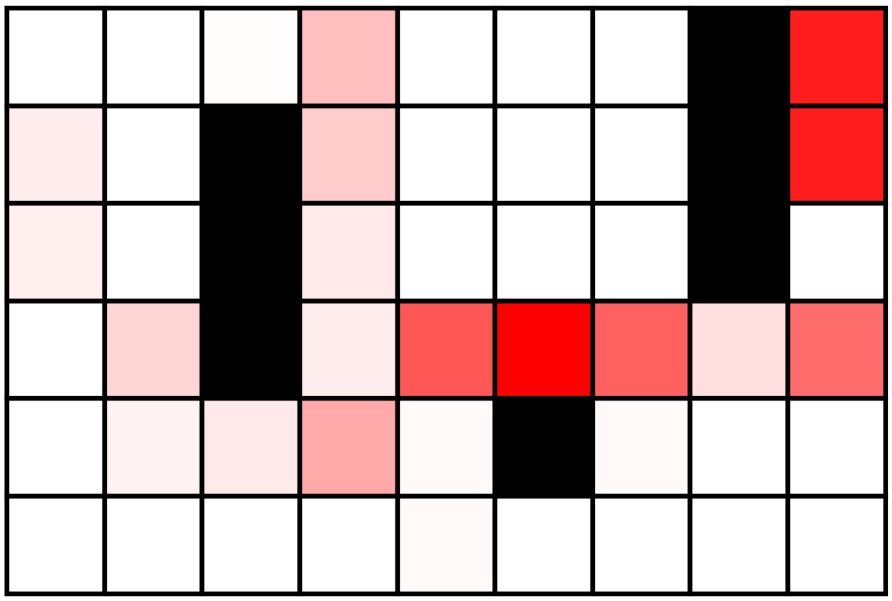

Figure 4. Location of the agent when it decides to replay during a sample run of the original GridWorld environment. Darker red indicates more replay activity. Given that replay is more advantageous in certain states than others, we correspondingly predict that agents will feel more fatigued in some states than others.

A Normative Lens to Understand Psychiatric Illnesses. Cognitve 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 (Montague, Dolan, Friston, & Dayan, 2012; Wang & Krystal, 2014; Huys, Maia, & Frank, 2016). Cognitive control is often considered to be disrupted in many psychiatric illnesses (J. D. Cohen & Servan-Schreiber, 1992; Braver, Barch, & Cohen, 1999), 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 (Mattar & Daw, 2018; Lieder, Griffiths, & Hsu, 2018) 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 (Kumaran et al., 2016; Gagne, Dayan, & Bishop, 2018).

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 (G. Cohen & Burke, 1993). 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., Corbit, Ostlund, & Balleine, 2002).

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 (Jadhav,

Rothschild, Roumis, & Frank, 2016; Shin, Tang, & Jadhav, 2019; McCormick, Barry, Jafarian, Barnes, & Maguire, 2020), 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 (Shenhav et al., 2013). 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.

These possibilities represent potentially valuable directions for future research concerning the mechanistic basis of fatigue. However, there is one phenomenon concerning the relation between performance and time on task that neither fatigue alone, nor the model as presented so far, can fully explain. We turn to that in the section that follows.

Part 2: Boredom

The EVB_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.

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 (Yerkes & Dodson, 1908). 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 (Csikszentmihalyi, 1997). 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 norepinephine (NE) function – a neuromodulator that is widely distributed throughout the brain. The Adaptive Gain Theory (AGT; Aston-Jones & Cohen, 2005) 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 Aston-Jones & Cohen, 2005 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, that otherwise might be viewed as 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 (Angela & Dayan, 2005). 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; Kane et al., 2017) and from the use of pupillometry as an indirect measure of LC function (Pajkossy, Szőllősi, Demeter, & Racsmány, 2017).

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.

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.

1. Optimal Participant-Task Fit. A survey by Fisher (1987) 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’ (Csikszentmihalyi, 1996, 1997; Jackson & Csikszentmihalyi, 1999; Perry, 1999; Nakamura & Csikszentmihalyi, 2014), the state of being completely immersed and focused in a task.

A recent set of experiments by Geana, Wilson, Daw, and Cohen (2016a) 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. (2020).

2. 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. (2016a) 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 (Fisher, 1987). In the extreme, boredom has been found to lead to a higher labor turnover (Wild & Hill, 1970; Geiwitz, 1966; Kishida, 1973).

Finally, Geana et al. (2016a) 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 (Berry & Fristedt, 1985) 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 for 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 (Schmidhuber, 1991; Oudeyer, Kaplan, & Hafner, 2007; R. C. Wilson, Shenhav, Straccia, & Cohen, 2019). In the section that follows, we generalize these formalizations and evaluate the extent to which it contributes to patterns of modulation of performance.

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 Part 1, providing a formal, integrated framework for investigating potential relationships between boredom and fatigue, as follows:

$$VOI(s_k, a_k) = Gain(s_k, a_k) \times Need(s_k, a_k) \tag{8}$$

in which

$$Gain(s_k, a_k) = \mathbb{E}\left[\sum_{a \in A}\Big[\pi_{new}(a|s_k) - \pi_{old}(a|s_k)\Big] q(s_k, a)\right] \tag{9}$$

$$= \int\limits_{r \sim R(s_k, a_k)} p(r)\left[\sum_{a \in A}\Big[\pi_{new}(a|s_k) - \pi_{old}(a|s_k)\Big] q_{new}(s_k, a)\right] dr \tag{10}$$

and

$$Need(s_k, a_k) = \sum_{i=1}^{\infty} \gamma^i \delta_{s_{t+i}, s_k} \tag{11}$$

$$= \gamma \sum_{s' \in S} T(s_k, a_k, s') M(s', s_k) \tag{12}$$

in which M is the successor representation (Dayan, 1993; Gershman et al., 2012). 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*. [5]

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: $\max_k VOI(s, a_k) - VOI(s, a_t)$[6].

[5] These 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

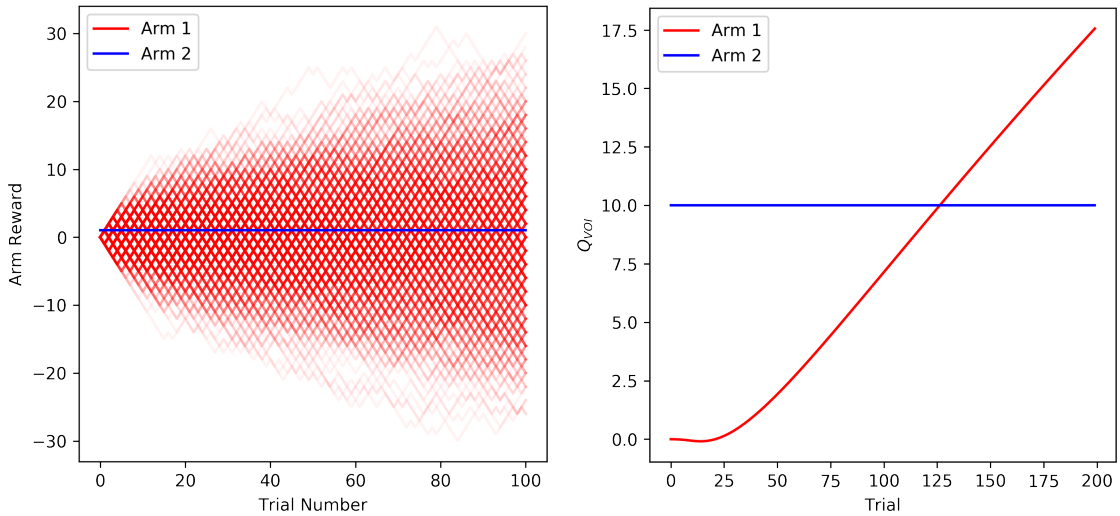

Figure 5. Sample bandit task illustrating the increase in Q_{VOI} with increased uncertainty. (Left) Red indicates sample reward trajectories of the first arm, the reward for which starts at zero and then increases or decreases by one every iteration. The blue arm 2 indicates an arm with value always 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 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

used for EVB by M&D, as in the previous section, and is sufficient for our purposes.

[6] An 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.

changes over 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 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 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 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 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

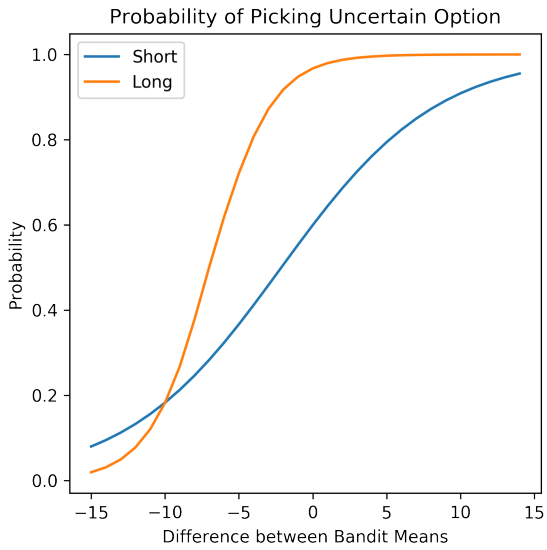

Figure 6. For various mean differences between the two bandit arms, we plot the softmax probability of choosing the option with higher uncertainty in the short horizon versus the long horizon. We reproduce the effect found in R. C. Wilson et al. (2014), in which it was shown that participants in longer 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 Need term. This can be examined, for instance, by manipulating the horizon in multi-armed bandit settings. For example, Figure 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 R. C. Wilson et al. (2014): participants explore more in games with longer horizons, when they have more opportunities to gain information about 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.

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 (1982) computational level, a complete theory would account for effects at all levels of analysis. As discussed before, Aston-Jones and Cohen (2005) 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, Kane et al. (2017) 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 (Schultz, 1998; Dabney et al., 2020).

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, that is evidently more directed toward individual alternatives, is curiosity. For instance, Dubey and Griffiths (2019) 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 Geana et al. (2016a)'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?

Part 3: Replaying to Explore

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. Part 1 modeled how internal computations — replay — can improve learning by propagating knowledge about rewards to distal states and actions. Part 2 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 Part 1) is more difficult than in bandit tasks (like Part 2) 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.

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 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 (Peng & Williams, 1993; Moore & Atkeson, 1993):

$$Gain(s_k, a_k) = \gamma^\tau \rho |R(s_k, a_k) + \gamma \max_a Q_{VOI}(s_{k+1}, a) - Q_{VOI}(s_k, a_k)| \tag{13}$$

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. [7] 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

$$(1 - \gamma^{\tau}) \sum_{a} \pi(a|s) Q(s, a) \tag{14}$$

Thus, the agent should replay the state-action pair with the highest EVB_C 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):

$$VOI(s) = U_0 \times (1 - e^{-kN_t(s)}) \tag{15}$$

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 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 7 demonstrates, an agent with this exploration bonus can use replay to achieve a high average reward rate.

[7] Underestimation arises because EVB_C 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.

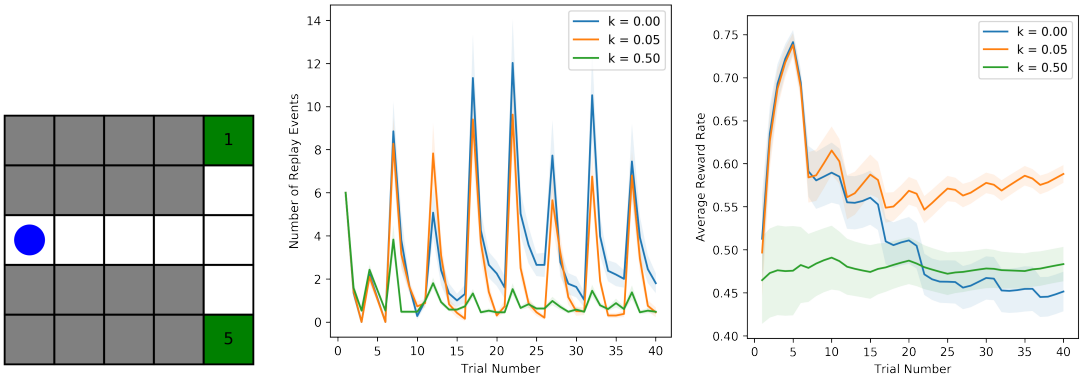

Figure 7. Simulations of an agent that replays and explores. The agent uses replay in order to propagate value of information through the agent's model of the environment. (Left) Gridworld T-maze environment. (Middle) Replay behavior for different hazard rates. (Right) Average reward rate for different hazard rates. Reward rate is greatest for a moderate hazard rate. All error bars indicate ± 1 SEM.

Discussion

It has sometimes been questioned whether performance decrements following long time-on-task are a result of fatigue or boredom (J. F. Mackworth, 1968; Pattyn, Neyt, Henderickx, & Soetens, 2008) Meanwhile, other work has not discriminated between these states, implicitly considering both as reflecting motivation and/or opportunity costs (Hockey, 2011; Kurzban et al., 2013). 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 prevalance in daily life (Smallwood & Schooler, 2006; Fox, Nijeboer, Solomonova, Domhoff, & Christoff, 2013; Fox & Christoff, 2014; Smallwood & Schooler, 2015; Christoff, Irving, Fox, Spreng, & Andrews-Hanna, 2016). Recent work has suggested that mind wandering can be goal-directed, and can facilitate creativity (Zedelius & Schooler, 2015; Williams et al., 2018; Fox & Christoff, 2018; Agnoli, Vanucci, Pelagatti, & Corazza, 2018). 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.

General Discussion

In this article we present a model that provides a formally rigorous, normative interpretation of the phenomena of fatigue and boredom associated with control-demanding tasks. This rests on the widely held assumption that the number of physical and mental tasks that can be performed at once is limited. This imples that engaging in one carries opportunity costs, which our models formalize in terms of the future value of replay for learning (signaled by fatigue) and information gathering through exploration (signaled by boredom). Both of these options involve an intertemporal tradeoff, in that they have the potential to earn greater rewards in the future at the cost of forestalling more immediate reward gathering. This account provides a mechanistic grounding for the intuitive idea that fatigue occurs when performing control-demanding tasks, especially ones that are more difficult (i.e., require more internal computation via replay to learn to perfect); and that boredom occurs when performing easy and repetitive tasks, especially for extended durations (i.e., increasing the likelihood that other, more remunerative opportunities have become available).

The theory predicts that fatigue should track the value of offline processing mechanisms such as hippocampal replay. One of the functions of replay is learning — that is, to facilitate the

transfer from controlled to automatic processing — and thus the value of this function is maximized when the agent is still control-dependent. As a result, our account predicts fatigue in control-dependent tasks but not once the agent develops automaticity.

Boredom arises in situations in which the likelihood increases that another more valuable task may be available, favoring the value of exploration. We formalized this exploratory value as the 'value of information' and suggested that it tracks the information content available when doing a task. Both easy and impossible tasks offer little information gain, and thus we expect these tasks to be the more boring. Boredom is minimized in 'flow' states, in which there is both a high exploitative and high exploratory value in the current task.

By providing a formal distinction between fatigue and boredom, our model may help guide future empirical work that addresses these phenomena. For example, we previously discussed the multi-hour N-back task from Blain et al. (2016), in which those in the harder task condition exhibited increased delay discounting impulsivity over the course of the experiment. We (concurring with the original authors) interpreted this behavior to reflect fatigue as opposed to boredom. Our computational framework specifies why. When opportunity costs are increasing during a control-dependent task, fatigue is responsible. Had it been boredom, we would have seen the opposite: those in the easier task condition would change their impulsivity more than those in the harder task. Of course, given how trivial the easy condition was, it is likely that those participants did become bored over time. Our claim is that this increase in boredom was not measured by delay discounting impulsivity. We would hypothesize, following the experiment from Geana et al. (2016a), that if the participants in the easy task condition were given a multi-armed bandit task, they would be more exploratory than those those in the harder task condition.

Lastly, we conjecture that one reason the line between fatigue and boredom can seem blurred is that replay and exploration can interact. Specifically, agents can use replay to propagate the value of information throughout their model of the environment, thus helping them plan in a way that includes exploration. While this interaction is suggested on purely computational

grounds, it raises the possibility that this interaction may be reflected in other forms of phenomenology, such as mind-wandering, offering the potential for a formally rigorous approach to interpreting those phenomena as well.

Limitations & Future Directions

Algorithmic Approximations. Parts 1 and 2 formalized the benefits of replaying and exploring in reinforcement learning environments, but the calculations may not always be feasible. For example, the Mattar and Daw (2018) model (and our version of it in Part 1) computes the value of all replay events before replaying the highest valued event. Such simulations allow us to expose the characteristics of optimal replay, but this is not viable (and not meant) as a realizable process-level account since the selection would take more computation than the computation being prioritized. More realizable, but more approximate, heuristics such as the prioritized sweeping formulation used in Part 3 have been proposed in the computer science (Peng & Williams, 1993; Moore & Atkeson, 1993; Schaul et al., 2015) and neuroscience (Momennejad et al., 2018) literatures, but it remains an open question as to how the brain ultimately computes these values. Similarly, the value of information metric relies on an integral which is intractable in most tasks. Agents may approximate this metric through heuristic algorithms such as UCB (Auer et al., 2002). An alternate approach might be to use learning rate (R. C. Wilson et al., 2019), a heuristic which has improved performance in machine learning environments (Schmidhuber, 1991; Şimşek & Barto, 2006).

Causal Disruptions of Fatigue. Although recent empirical studies have begun to test the extent to which boredom plays a causal role in signaling the value of exploration (e.g., Geana et al., 2016a; Geana, Wilson, Daw, & Cohen, 2016b), there has not yet been a direct test of the extent to which cognitive fatigue signals the value of replay. Part of the problem lies in creating an adequate control that would rule out metabolic resource theories. For example, simply stopping participants from taking a break in order to prevent replay would not be an informative manipulation, because a metabolic resource theory would also predict these participants to be

fatigued. A more direct test would involve disrupting mechanisms of offline processing in humans (e.g., using transcranial magnetic stimulation, or direct current stimulation), similar to the disruption of sharp wave ripples in rodents (Girardeau, Benchenane, Wiener, Buzsáki, & Zugaro, 2009; Jadhav, Kemere, German, & Frank, 2012), and measuring its impact on performance, reports of fatigue, and inclinations to rest. This represents an important direction for future research.

Conclusion

The opportunity costs associated with cognitive control exhibit complex temporal dynamics. In this article, we proposed two mechanisms that give rise to these costs: mental simulation (replay) and exploration, that are tracked by fatigue and boredom, respectively. Both reflect an intertemporal choice agents must make in the pursuit of reward maximization. We explained how the independent dynamics of these mechanisms may unify a range of disparate findings in the literature on cognitive control, and proposed that they might interact in a novel way, enabling agents to plan to explore. More generally, they help place cognitive control in the context of approaches, such as bounded rationality and resource rationality (Simon, 1972; Howes, Lewis, & Vera, 2009; Shenhav et al., 2013; Lieder & Griffiths, 2020), that assume agents optimize their utility functions based on a cost-benefit analysis, constrained by the resources and time they have available for computation and action. We hope this provides a useful foundation for future work involving both experimental tests and refinement of theory.

Acknowledgements

We thank James Antony, Ishita Dasgupta, and Rachit Dubey for helpful comments on previous drafts. This work is supported by the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. M.A. is supported by the National Defense Science and Engineering Graduate Fellowship Program.

References

Agnoli, S., Vanucci, M., Pelagatti, C., & Corazza, G. E. (2018). Exploring the link between mind wandering, mindfulness, and creativity: A multidimensional approach. *Creativity Research Journal*, *30*(1), 41–53.

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem situations. *Psychological review*, *94*(2), 192.

Angela, J. Y., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692.

Antony, J. W., Hartshorne, T. H., Pomeroy, K., Gureckis, T. M., Hasson, U., McDougle, S. D., & Norman, K. A. (2020). Behavioral, physiological, and neural signatures of surprise during naturalistic sports viewing. *BioRxiv*.

Arai, T. (1912). *Mental fatigue* (No. 54). Teachers College, Columbia University.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, *28*, 403–450.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, *47*(2-3), 235–256.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of personality and social psychology*, *74*(5), 1252.

Baumeister, R. F., & Vohs, K. D. (2007). Self-regulation, ego depletion, and motivation. *Social and personality psychology compass*, *1*(1), 115–128.

Bergum, B. O., & Lehr, D. J. (1962). Vigilance performance as a function of interpolated rest. *Journal of Applied Psychology*, *46*(6), 425.

Berry, D. A., & Fristedt, B. (1985). Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, *5*(71-87), 7–7.

Blain, B., Hollard, G., & Pessiglione, M. (2016). Neural mechanisms underlying the impact of daylong cognitive work on economic decisions. *Proceedings of the National Academy of*

Sciences, *113*(25), 6967–6972.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *The Journal of Educational Research*, *74*(4), 245–248.

Braver, T. S., Barch, D. M., & Cohen, J. D. (1999). Cognition and control in schizophrenia: a computational model of dopamine and prefrontal function. *Biological psychiatry*, *46*(3), 312–328.

Carr, M. F., Jadhav, S. P., & Frank, L. M. (2011). Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nature neuroscience*, *14*(2), 147.

Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). 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.

Christoff, K., Irving, Z. C., Fox, K. C., Spreng, R. N., & Andrews-Hanna, J. R. (2016). Mind-wandering as spontaneous thought: a dynamic framework. *Nature Reviews Neuroscience*, *17*(11), 718.

Cohen, G., & Burke, D. M. (1993). Memory for proper names: A review. *Memory*, *1*(4), 249–263.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychological review*, *97*(3), 332.

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

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychological review*, *99*(1), 45.

Corbit, L. H., Ostlund, S. B., & Balleine, B. W. (2002). 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.

Csikszentmihalyi, M. (1996). *Creativity: The work and lives of 91 eminent people*. HarperCollins.

Csikszentmihalyi, M. (1997). Flow and the psychology of discovery and invention. *HarperPerennial, New York*, *39*.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, 1–5.

Dasgupta, I., Smith, K. A., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2018). Learning to act by integrating mental simulations and physical experiments. *BioRxiv*, 321497.

Davidson, T. J., Kloosterman, F., & Wilson, M. A. (2009). Hippocampal replay of extended experience. *Neuron*, *63*(4), 497–507.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*(12), 1704.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624.

Diba, K., & Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature neuroscience*, *10*(10), 1241.

Dodge, R. (1917). The laws of relative fatigue. *Psychological Review*, *24*(2), 89.

Donovan, J. J., & Radosevich, D. J. (1999). 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.

Dora, J., van Hooff, M., Geurts, S., Kompier, M., & Bijleveld, E. (2019). The effect of opportunity costs on mental fatigue in labor/leisure tradeoffs.

Dubey, R., & Griffiths, T. (2019). Reconciling novelty and complexity through a rational analysis

of curiosity.

Eldar, E., Lièvre, G., Dayan, P., & Dolan, R. J. (2020). The roles of online and offline replay in planning. *BioRxiv*.

Fisher, C. D. (1987). *Boredom: Construct, causes and consequences* (Tech. Rep.). TEXAS A AND M UNIV COLLEGE STATION DEPT OF MANAGEMENT.

Foster, D. J., & Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, *440*(7084), 680.

Fox, K. C., & Christoff, K. (2014). Metacognitive facilitation of spontaneous thought processes: when metacognition helps the wandering mind find its way. In *The cognitive neuroscience of metacognition* (pp. 293–319). Springer.

Fox, K. C., & Christoff, K. (2018). *The oxford handbook of spontaneous thought: Mind-wandering, creativity, and dreaming*. Oxford University Press.

Fox, K. C., Nijeboer, S., Solomonova, E., Domhoff, G. W., & Christoff, K. (2013). Dreaming as mind wandering: evidence from functional neuroimaging and first-person content reports. *Frontiers in human neuroscience*, *7*, 412.

Gagne, C., Dayan, P., & Bishop, S. J. (2018). 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.

Geana, A., Wilson, R., Daw, N. D., & Cohen, J. D. (2016a). Boredom, information-seeking and exploration. In *Cogsci.*

Geana, A., Wilson, R., Daw, N. D., & Cohen, J. D. (2016b). Information-seeking, learning and the marginal value theorem: A normative approach to adaptive exploration. In *Cogsci.*

Geiwitz, P. J. (1966). Structure of boredom. *Journal of personality and social psychology*, *3*(5), 592.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential

decision making: A tale of two systems. *Journal of Experimental Psychology: General*, *143*(1), 182.

Gershman, S. J., Moore, C. D., Todd, M. T., Norman, K. A., & Sederberg, P. B. (2012). The successor representation and temporal context. *Neural Computation*, *24*(6), 1553–1568.

Gibson, E. L. (2007). Carbohydrates and mental function: feeding or impeding the brain? *Nutrition Bulletin*, *32*, 71–83.

Girardeau, G., Benchenane, K., Wiener, S. I., Buzsáki, G., & Zugaro, M. B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. *Nature neuroscience*, *12*(10), 1222.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(2), 148–164.

Gupta, A. S., van der Meer, M. A., Touretzky, D. S., & Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. *Neuron*, *65*(5), 695–705.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . others (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573.

Healy, A. F., Kole, J. A., Buck-Gengler, C. J., & Bourne, L. E. (2004). Effects of prolonged work on data entry speed and accuracy. *Journal of Experimental Psychology: Applied*, *10*(3), 188.

Helton, W. S., & Russell, P. N. (2015). Rest is best: The role of rest and task interruptions on vigilance. *Cognition*, *134*, 165–173.

Helton, W. S., & Russell, P. N. (2017). Rest is still best: The role of the qualitative and quantitative load of interruptions on vigilance. *Human factors*, *59*(1), 91–100.

Hockey, G. R. J. (2011). A motivational control theory of cognitive fatigue. *Cognitive fatigue: Multidisciplinary perspectives on current research and future applications*, 167–87.

Howes, A., Lewis, R. L., & Vera, A. (2009). Rational adaptation under task and processing constraints: Implications for testing theories of cognition and action. *Psychological review*,

116(4), 717.

Huxtable, Z. L., White, M. H., & McCartor, M. A. (1946). A re-performance and re-interpretation of the arai experiment in mental fatigue with three subjects. *Psychological Monographs*, *59*(5), 1–52.

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, *19*(3), 404.

Inzlicht, M., Schmeichel, B. J., & Macrae, C. N. (2014). Why self-control seems (but may not be) limited. *Trends in cognitive sciences*, *18*(3), 127–133.

Izawa, C. (1971). Massed and spaced practice in paired-associate learning: List versus item distributions. *Journal of Experimental Psychology*, *89*(1), 10.

Jackson, S. A., & Csikszentmihalyi, M. (1999). *Flow in sports*. Human Kinetics.

Jadhav, S. P., Kemere, C., German, P. W., & Frank, L. M. (2012). Awake hippocampal sharp-wave ripples support spatial memory. *Science*, *336*(6087), 1454–1458.

Jadhav, S. P., Rothschild, G., Roumis, D. K., & Frank, L. M. (2016). Coordinated excitation and inhibition of prefrontal ensembles during awake hippocampal sharp-wave ripple events. *Neuron*, *90*(1), 113–127.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, *4*, 237–285.

Kane, G. A., Vazey, E. M., Wilson, R. C., Shenhav, A., Daw, N. D., Aston-Jones, G., & Cohen, J. D. (2017). Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience*, *17*(6), 1073–1083.

Karlsson, M. P., & Frank, L. M. (2009). Awake replay of remote experiences in the hippocampus. *Nature neuroscience*, *12*(7), 913.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS computational biology*, *7*(5), e1002055.

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). 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.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, *55*(4), 352.

Kishida, K. (1973). Temporal change of subsidiary behavior in monotonous work. *Journal of Human Ergology*, *2*(1), 75–89.

Kool, W., & Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology: General*.

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLoS computational biology*, *12*(8), e1005090.

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological science*, *28*(9), 1321–1333.

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, *20*(7), 512–534.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and brain sciences*, *36*(6), 661–679.

Lee, A. K., & Wilson, M. A. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, *36*(6), 1183–1194.

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

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, *125*(1), 1.

Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. (2019). Human replay spontaneously reorganizes experience. *Cell*, *178*(3), 640–652.

Lorist, M. M., Boksem, M. A., & Ridderinkhof, K. R. (2005). Impaired cognitive control and

reduced cingulate activity during mental fatigue. *Cognitive Brain Research*, *24*(2), 199–205.

Louie, K., & Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, *29*(1), 145–156.

Mackworth, J. F. (1968). Vigilance, arousal, and habituation. *Psychological review*, *75*(4), 308.

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, *1*(1), 6–21.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information.

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience*, *21*(11), 1609.

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

McCormick, C., Barry, D. N., Jafarian, A., Barnes, G. R., & Maguire, E. A. (2020). vmpfc drives hippocampal processing during autobiographical memory recall regardless of remoteness. *bioRxiv*.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., … Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191.

Messier, C. (2004). Glucose improvement of memory: a review. *European journal of pharmacology*, *490*(1-3), 33–57.

Metcalfe, J., & Xu, J. (2016). People mind wander more during massed than spaced inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(6), 978.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes

and multiple-task performance: Part i. basic mechanisms. *Psychological review*, *104*(1), 3.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … others (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529–533.

Molden, D. C., Hui, C. M., Scholer, A. A., Meier, B. P., Noreen, E. E., D'Agostino, P. R., & Martin, V. (2012). Motivational versus metabolic effects of carbohydrates on self-control. *Psychological science*, *23*(10), 1137–1144.

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *Elife*, *7*, e32548.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of neuroscience*, *16*(5), 1936–1947.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, *16*(1), 72–80.

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

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, *31*, 69–89.

Musslick, S., Dey, B., Özcimder, K., Patwary, M. M. A., Willke, T. L., & Cohen, J. D. (2016). Controlled vs. automatic processing: A graph-theoretic approach to the analysis of serial vs. parallel processing in neural network architectures. In *Cogsci.*

Nádasdy, Z., Hirase, H., Czurkó, A., Csicsvari, J., & Buzsáki, G. (1999). Replay and time compression of recurring spike sequences in the hippocampus. *Journal of Neuroscience*, *19*(21), 9497–9507.

Nakamura, J., & Csikszentmihalyi, M. (2014). The concept of flow. In *Flow and the foundations of positive psychology* (pp. 239–263). Springer.

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological review*, *86*(3), 214.

Newberg, A. B., Wang, J., Rao, H., Swanson, R. L., Wintering, N., Karp, J. S., … Detre, J. A. (2005). Concurrent cbf and cmrglc changes during human brain activation by combined fmri–pet scanning. *Neuroimage*, *28*(2), 500–506.

Nicholson, W., & Snyder, C. M. (2012). *Microeconomic theory: Basic principles and extensions*. Nelson Education.

Niyogi, R. K., Breton, Y.-A., Solomon, R. B., Conover, K., Shizgal, P., & Dayan, P. (2014). Optimal indolence: a normative microscopic approach to work and leisure. *Journal of The Royal Society Interface*, *11*(91), 20130969.

Niyogi, R. K., Shizgal, P., & Dayan, P. (2014). Some work and some play: microscopic and macroscopic approaches to labor and leisure. *PLoS computational biology*, *10*(12), e1003894.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain research*.

Ólafsdóttir, H. F., Bush, D., & Barry, C. (2018). The role of hippocampal replay in memory and planning. *Current Biology*, *28*(1), R37–R50.

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

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

Oudeyer, P.-Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, *11*(2), 265–286.

Pajkossy, P., Szőllősi, Á., Demeter, G., & Racsmány, M. (2017). Tonic noradrenergic activity modulates explorative behavior and attentional set shifting: Evidence from pupillometry and gaze pattern analysis. *Psychophysiology*, *54*(12), 1839–1854.

Pattyn, N., Neyt, X., Henderickx, D., & Soetens, E. (2008). Psychophysiological investigation of

vigilance decrement: boredom or cognitive fatigue? *Physiology & behavior*, *93*(1-2), 369–378.

Peng, J., & Williams, R. J. (1993). Efficient learning and planning within the dyna framework. *Adaptive Behavior*, *1*(4), 437–454.

Perry, S. K. (1999). *Writing in flow: Keys to enhanced creativity*. Writers Digest Books.

Pfeiffer, B. E., & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, *497*(7447), 74.

Posner, M., & Snyder, C. R. (1975). Facilitation and inhibition. *Attention and performance*.

Raichle, M. E., & Mintun, M. A. (2006). Brain work and brain imaging. *Annu. Rev. Neurosci.*, *29*, 449–476.

Randles, D., Harlow, I., & Inzlicht, M. (2017). A pre-registered naturalistic observation of within domain mental fatigue and domain-general depletion of self-control. *PloS one*, *12*(9), e0182980.

Rea, C. P., & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning: Journal of Practical Research & Applications*.

Ross, H. A., Russell, P. N., & Helton, W. S. (2014). Effects of breaks and goal switches on the vigilance decrement. *Experimental brain research*, *232*(6), 1729–1737.

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological review*, *115*(1), 101.

Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C., & Norman, K. A. (2018). Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nature communications*, *9*(1), 3920.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). 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.

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). Prioritized experience replay. *arXiv preprint arXiv:1511.05952*.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological methods*, *17*(4), 551.

Schmidhuber, J. (1991). 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* (pp. 222–227).

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. detection, search, and attention. *Psychological review*, *84*(1), 1.

Schuck, N. W., & Niv, Y. (2019). Sequential replay of nonspatial task states in the human hippocampus. *Science*, *364*(6447), eaaw5181.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, *80*(1), 1–27.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Schulz, E., Bhui, R., Love, B. C., Brier, B., Todd, M. T., & Gershman, S. J. (2019). Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences*, *116*(28), 13903–13908.

Sezener, C. E., Dezfouli, A., & Keramati, M. (2019). Optimizing the depth and the direction of prospective planning using information values. *PLoS computational biology*, *15*(3), e1006827.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: Ii. perceptual learning, automatic attending and a general theory. *Psychological review*, *84*(2), 127.

Shin, J. D., Tang, W., & Jadhav, S. P. (2019). Dynamics of awake hippocampal-prefrontal replay

for spatial learning and memory-guided decision making. *Neuron*.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and organization*, *1*(1), 161–176.

Şimşek, Ö., & Barto, A. G. (2006). An intrinsic reward mechanism for efficient exploration. In *Proceedings of the 23rd international conference on machine learning* (pp. 833–840).

Smallwood, J., & Schooler, J. W. (2006). The restless mind. *Psychological bulletin*, *132*(6), 946.

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: empirically navigating the stream of consciousness. *Annual review of psychology*, *66*, 487–518.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychological review*, *119*(1), 120.

Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, *2*(4), 160–163.

Sutton, R. S., Barto, A. G., et al. (1998). *Introduction to reinforcement learning* (Vol. 2) (No. 4). MIT press Cambridge.

Tambini, A., Ketz, N., & Davachi, L. (2010). Enhanced brain correlations during rest are related to memory for recent experiences. *Neuron*, *65*(2), 280–290.

Thorndike, E. (1900). Mental fatigue. i. *Psychological Review*, *7*(6), 547.

Thorndike, E. L. (1912). The curve of work. *Psychological Review*, *19*(3), 165.

Tononi, G., & Cirelli, C. (2003). Sleep and synaptic homeostasis: a hypothesis. *Brain research bulletin*, *62*(2), 143–150.

Tononi, G., & Cirelli, C. (2006). Sleep function and synaptic homeostasis. *Sleep medicine reviews*, *10*(1), 49–62.

Tononi, G., & Cirelli, C. (2014). Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration. *Neuron*, *81*(1), 12–34.

Van der Linden, D., Frese, M., & Meijman, T. F. (2003). Mental fatigue and the control of cognitive processes: effects on perseveration and planning. *Acta psychologica*, *113*(1), 45–65.

Van Der Meer, M. A., & Redish, A. D. (2009). Covert expectation-of-reward in rat ventral striatum at decision points. *Frontiers in integrative neuroscience*, *3*, 1.

van de Ven, G. M., Trouche, S., McNamara, C. G., Allen, K., & Dupret, D. (2016). Hippocampal offline reactivation consolidates recently formed cell assembly patterns during sharp wave-ripples. *Neuron*, *92*(5), 968–974.

Wamsley, E. J. (2019). Memory consolidation during waking rest. *Trends in cognitive sciences*, *23*(3), 171–173.

Wang, X.-J., & Krystal, J. H. (2014). Computational psychiatry. *Neuron*, *84*(3), 638–654.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human factors*, *50*(3), 433–441.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3-4), 279–292.

Wikenheiser, A. M., & Redish, A. D. (2015). Hippocampal theta sequences reflect current goals. *Nature neuroscience*, *18*(2), 289.

Wild, R., & Hill, A. (1970). *Women in the factory: A study of job satisfaction and labour turnover*. Institute of Personnel Management.

Williams, K. J., Lee, K. E., Hartig, T., Sargent, L. D., Williams, N. S., & Johnson, K. A. (2018). Conceptualising creativity benefits of nature experience: Attention restoration and mind wandering as complementary processes. *Journal of Environmental Psychology*, *59*, 36–45.

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, *265*(5172), 676–679.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.

Wilson, R. C., Shenhav, A., Straccia, M., & Cohen, J. D. (2019). The eighty five percent rule for optimal learning. *Nature communications*, *10*(1), 1–9.

Wimmer, G. E., Liu, Y., Vehar, N., Behrens, T. E., & Dolan, R. J. (2019). Episodic memory retrieval is supported by rapid replay of episode content. *bioRxiv*, 758185.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology and psychology*, *18*(5), 459–482.

Zedelius, C. M., & Schooler, J. W. (2015). Mind wandering “ahas” versus mindful reasoning: alternative routes to creative solutions. *Frontiers in Psychology*, *6*, 834.

Appendix

Methods

Part 1: Cognitive Fatigue. 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:

$$\pi(a|s) = \frac{e^{\beta Q(s,a)}}{\sum_{a'} e^{\beta Q(s,a')}} \tag{16}$$

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. 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 EVB_C is computed. If the value is greater than zero, the agent replays $\arg\max EVB_C$. 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 and Figures 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 $\sigma = 5$.

Part 2: Boredom. The right panel of Figure 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 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.

Part 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. 15, 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.