Psychological Review

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

Mayank Agrawal, Marcelo G. Mattar, Jonathan D. Cohen, and Nathaniel D. Daw

CITATION

Agrawal, M., Mattar, M. G., Cohen, J. D., & Daw, N. D. (2021, August 12). The Temporal Dynamics of Opportunity Costs: A Normative Account of Cognitive Fatigue and Boredom. *Psychological Review*. Advance online publication. http://dx.doi.org/10.1037/rev0000309

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

ISSN: 0033-295X

https://doi.org/10.1037/rev0000309

THEORETICAL NOTE

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

Mayank Agrawal[1], Marcelo G. Mattar[2], Jonathan D. Cohen[1], and Nathaniel D. Daw[1]
[1] Department of Psychology, Princeton Neuroscience Institute, Princeton University
[2] Department of Cognitive Science, University of California San Diego

Cognitive fatigue and boredom are two phenomenological states that reflect overt task disengagement. In this article, we present a rational analysis of the temporal structure of controlled behavior, which provides a formal account of these phenomena. We suggest that in controlling behavior, the brain faces competing behavioral and computational imperatives, and must balance them by tracking their opportunity costs over time. We use this analysis to flesh out previous suggestions that feelings associated with subjective effort, like cognitive fatigue and boredom, are the phenomenological counterparts of these opportunity cost measures, instead of reflecting the depletion of resources as has often been assumed. Specifically, we propose that both fatigue and boredom reflect the competing value of particular options that require foregoing immediate reward but can improve future performance: Fatigue reflects the value of offline computation (internal to the organism) to improve future decisions, while boredom signals the value of exploration (external in the world). We demonstrate that these accounts provide a mechanistically explicit and parsimonious account for a wide array of findings related to cognitive control, integrating and reimagining them under a single, formally rigorous framework.

Keywords: reinforcement learning, hippocampal replay, explore-exploit, cognitive control

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 (Anderson, 1987; Posner & Snyder, 1975; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). 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 (Cohen et al., 1990; Daw et al., 2005; Sutton & Barto, 1998).

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 2 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 14 straight hours of practice is comparable to that at the end of the same amount of practice carried out periodically over 2 weeks.[1] This is unlikely to be true (Arai, 1912; Haager et al., 2018; Healy et al., 2004; Huxtable et al., 1946; Lorist et al., 2005; Mackworth, 1948; Van der Linden et al., 2003; Warm et al., 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 resource (e.g., metabolic; Baumeister & Vohs, 2007; Baumeister et al., 1998), by analogy to the case of physical fatigue. In contrast, Kurzban et al. (2013) proposed that

Mayank Agrawal https://orcid.org/0000-0003-3541-1427
Marcelo G. Mattar https://orcid.org/0000-0003-3303-2490
Nathaniel D. Daw https://orcid.org/0000-0001-5029-1430

Jonathan D. Cohen and Nathaniel D. Daw, listed alphabetically, made equal and complementary contributions to the guidance and support of the work reported in this article.

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. Mayank Agrawal is supported by the National Defense Science and Engineering Graduate Fellowship Program.

A preliminary version of the model in Part 1 was presented at the 2019 Computational Cognitive Neuroscience (CCN) Conference.

Correspondence concerning this article should be addressed to Mayank Agrawal, Department of Psychology, Princeton Neuroscience Institute, Princeton University, Washington Rd, Princeton, NJ 08540, United States. Email: mayank.agrawal@princeton.edu

---

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

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 (Bench & Lench, 2013; Hockey, 2013; Inzlicht et al., 2014). 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 as follows: (a) It must define the nature of the alternative behaviors that give rise to the opportunity cost(s); and (b) 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, that can be used for learning. 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, here playing out via competition between different classes of overt action: Information seeking versus exploitative. We offer a model that enables agents to navigate the explore–exploit tradeoff, expose its analogy to the case of replay, and demonstrate how the temporal dynamics of uncertainty lead agents to oscillate between different tasks. Finally, 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.

Figure 1
Types of Task Engagement and Disengagement

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

Note. This article partitions the constraints associated with the duration of cognitive control into two dimensions: action versus replay and exploration versus exploitation. Part 1 investigates fatigue, which we propose to be a signal for the value of replay while Part 2 investigates boredom, which we propose to be a bias toward 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.

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

Background

Rational Models of Cognitive Control

Cognitive fatigue has long been studied in psychology (Dodge, 1917; Thorndike, 1900), with one influential account, “ego depletion,” suggesting that it reflects the consumption and subsequent diminution of a metabolic resource (Baumeister & Vohs, 2007; Baumeister et al., 1998). 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 et al., 2015; Hagger et al., 2016; Randles et al., 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 (Gibson, 2007; Messier, 2004; Molden et al., 2012; Raichle & Mintun, 2006; 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 (Kurzban et al., 2013; Lieder & Griffiths, 2020; Shenhav et al., 2013) propose that performance variation arises from the rational balancing of the costs versus benefits of different control strategies, rather than biological resource limitations. Although these accounts vary as to how they operationalize control (and thus what ultimately makes it costly or limited), the cost–benefit framing implies that performance decrements due to fatigue or boredom can be countered with incentives, shifting the tradeoff. The current theory instantiates a rational model to explain fatigue and boredom. Agents' actions in our model span two dimensions: Physical versus mental and exploratory versus exploitative (Figure 1). Because an agent can usually perform only 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 (Daw et al., 2005; Sutton & Barto, 1998) offers an integrative computational framework in which to implement a rational agent. Sequential decision problems in reinforcement learning settings are often modeled through a Markov decision process (MDP), a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$, in which $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\mathcal{R}(s)$ is the reward received in state s, $\mathcal{P}(s, a, s')$ is the probability of transitioning to from state s to state s' using action a, and γ is the discount factor. The policy $\pi : \mathcal{S} \mapsto \mathcal{A}$ determines with what probability an agent should perform action a when in state s.

Model-Free Versus Model-Based Learning

Two main classes of algorithms have emerged in reinforcement learning: Model-free and model-based learning (Daw et al., 2005; Sutton & Barto, 1998). 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 et al., 1996; Schultz, 1998; Schultz et al., 1997). 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:

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

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 Equation 1 teach the agent about local rewards and costs, but working out their consequences for longer-run action–outcome relationships requires additional mental simulation (or many more experiential updates). Importantly, such mental simulation can "teach" the agent how to make better choices in the future, but without actually collecting new information from the environment—instead, by discovering the consequences implicit in the information already known. 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 et al., 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; Keramati et al., 2016; Kool et al., 2016, 2017; Lee et al., 2014; Otto, Gershman, et al., 2013; Otto, Raio, et al., 2013; Sezener et al., 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 Equation 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 the hippocampus. This theory will be the basis of our analysis in Part 1, where we suggest fatigue tracks the value of such replay.

Exploration Versus 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 (Cohen et al., 2007; Daw et al., 2006; Gershman, 2018; Kaelbling et al., 1996; Mehlhorn et al., 2015; Schulz et al., 2019; Sutton & Barto, 1998; Wilson et al., 2014).

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, versus its expected long-run value including the improvement due to learning. For our purposes, we can generically express this as the sum of a baseline value Q and an additional increment, called the value of information (VOI):

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

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 the 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 et al., 2002), which proposes the VOI in a multi-armed bandit setting to be:

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

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 increasing boredom reflects a bias toward 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; Müller & Apps, 2019), with Edward Thorndike defining fatigue as "that diminution in efficiency *which rest can cure* (emphasis ours)" over 100 years ago (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; Helton & Russell, 2015, 2017; Ross et al., 2014). For example, Helton and Russell (2015) demonstrated the benefit of rest by having participants carry out a vigilance task (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 postinterruption 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 were predicted by the degree to which it involved a task that was distinct from (i.e., did not overlap) with the vigilance task: Those in the rest condition performed the best postinterruption, 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 versus rest reflects a valuation of mental effort and rest as nonsubstitutable "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 1 hr, 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 et al. (2016) conducted a study over the span of 6 hr. Participants performed either the "easy" tasks of a 1-back and 1-switch[2] or the "hard" tasks of a 3-back and 12-switch. Every 30 min, 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

---

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

N-switch) was comparable across groups and consistent throughout the experiment, participants in the "hard" condition made increasingly more impulsive choices (i.e., discounted more heavily) over the course of the experiment, whereas those in the "easy" condition did not show this effect. Assuming that increased impulsivity reflects fatigue, the results from this experiment suggest that participants in the "hard" condition were more fatigued than those in the "easy" condition.

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 (Carr et al., 2011; McClelland et al., 1995; Ólafsdóttir et al., 2018; Tambini et al., 2010; Wamsley, 2019). This suggests a grounding for the benefits of wakeful rest—that is, 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 the hippocampus called "place cells" are famously tuned to spatial locations, that is they tend to respond when the organism is in a certain location (Moser et al., 2008; O'Keefe & Dostrovsky, 1971). 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 (Davidson et al., 2009; Diba & Buzsáki, 2007; Foster & Wilson, 2006; Gupta et al., 2010; Karlsson & Frank, 2009). Replay events are commonly observed during epochs of high-frequency oscillatory activity in the hippocampus known as "sharp-wave ripples." When compared with the spatial locations represented by the place cells, the replayed sequential patterns often correspond to spatial trajectories—both experienced and novel—in the animal's physical environment (Louie & Wilson, 2001; Nádasdy et al., 1999; Wilson & McNaughton, 1994; 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 (Eldar et al., 2020; Gershman et al., 2014; Liu et al., 2019; Liu et al., 2020; Momennejad et al., 2018; Schapiro et al., 2018; Schuck & Niv, 2019; Wimmer et al., 2019).

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, the 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 electroencephalogram signal). 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 (Daw et al., 2005; Sutton & Barto, 1998). 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). This is useful, in turn, because experience is actually playing two roles in an algorithm like Q-learning. It is both interacting with the world to gather information about how a task works, for example, the location of rewards, but also propagating that information along experienced trajectories to work out its consequences for distal actions. The latter function (though not the former) can also be accomplished by mental simulation. For instance, even once you know the rules of chess completely, it takes further computation to elaborate their consequences for the best moves in particular situations. If this simulated experience is faster than physical experience and/or selected through a priority metric (Moore & Atkeson, 1993; Peng & Williams, 1993), the agent can converge to the optimal policy quicker, and thus increase the future reward, as opposed to relying exclusively on physical experience.

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

$$\begin{aligned}
&\mathrm{EVB}(s, s_k, a_k) \\
&\quad = \mathop{\mathbb{E}}_{\pi_{\text{new}}}\left[\sum_{i=0}^{\infty} \gamma^i R_{t+i} | S_t = s\right] - \mathop{\mathbb{E}}_{\pi_{\text{old}}}\left[\sum_{i=0}^{\infty} \gamma^i R_{t+i} | S_t = s\right], \qquad (4) \\
&\quad = \mathrm{Gain}(s_k, a_k) \times \mathrm{Need}(s, s_k). \qquad (5)
\end{aligned}$$

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

---

[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).

$$\mathrm{Gain}(s_k, a_k) = \sum_{a\in A} Q_{\pi_{\mathrm{new}}}(s_k, a)\pi_{\mathrm{new}}(a|s_k) - \sum_{a\in A} Q_{\pi_{\mathrm{new}}}(s_k, a)\pi_{\mathrm{old}}(a|s_k). \tag{6}$$

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, versus following the status quo policy π_{old}. The *Need* term corresponds to the expected number of (delay discounted) future visits to that state:

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

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. This in turn can be obtained from the successor representation M (Dayan, 1993; Gershman et al., 2012), estimated as $\hat{M}$, for the state pair (s, s_k).

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

Expected Value of Backup With Cost (EVB_C)

While the M&D model provides a rationale for *which* experiences an agent should replay (Moore & Atkeson, 1993; Peng & Williams, 1993; Schaul et al., 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 to 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, the Expected Value of Backup with Cost, $\mathrm{EVB}_C(s, s_k, a_k)$, is the expected increase in reward resulting from replaying the state-action pair (s_k, a_k) while in state s and executing the corresponding Bellman backup (i.e., temporal difference learning update, as in Equation 1 but for a simulated rather than experienced step), minus the amount of reward lost due to the time it takes to replay that state-action pair.

$$\mathrm{EVB}_C(s, s_k, a_k) = \gamma^{\tau}\mathrm{EVB}(s, s_k, a_k) - (1 - \gamma^{\tau})\sum_{a\in A} Q_{\pi_{\mathrm{old}}}(s, a)\pi_{\mathrm{old}}(a|s), \tag{8}$$

Here, the first term corresponds to the EVB discounted by τ which is the ratio of time it takes to replay versus act.[4] This discounting indicates that the benefits of replay can only be accrued after the time required to replay, τ, has elapsed. The second term is the reward lost due to the time it takes to replay: $\sum_{a\in A} Q_{\pi_{\mathrm{old}}}(s, a)\pi_{\mathrm{old}}(a|s)$ is the expected discounted future reward that would be available if the agent started acting immediately, and $\gamma^{\tau}\sum_{a\in A} Q_{\pi_{\mathrm{old}}}(s, a)\pi_{\mathrm{old}}(a|s)$ is the same quantity adjusted by the passage of τ. Their subtraction—that is, $(1 - \gamma^{\tau})\sum_{a\in A} Q_{\pi_{\mathrm{old}}}(s, a)\pi_{\mathrm{old}}(a|s)$—is thus the reward forgone due to the time it takes to replay. The derivation of this result can be found in the Appendix.

Hippocampal Replay as the Value of Leisure

The M&D model, and our subsequent EVB_C extension, provides a quantitative value to "rest" if the agent is engaging in replay during these rest states. Defining EVB_C^* as $\max \mathrm{EVB}_C(s, \cdot, \cdot)$, a rational agent should replay the most valuable location, $\arg\max \mathrm{EVB}_C(s, \cdot, \cdot)$, as long as its value $\mathrm{EVB}_C^* > 0$. Hence, if EVB_C^* is positive, replaying is more valuable than acting, a situation which we propose is subjectively sensed as fatigue. If EVB_C^* is negative acting is more valuable than replaying, and thus the agent should physically act instead of resting. Thus, the agent is optimizing the intertemporal tradeoff between acting (providing a more immediate opportunity for reward) and replaying (providing an opportunity for greater but later reward). This insight may help to rationalize the labor and leisure tradeoff that has been described for cognitive control (Dora et al., 2019; Inzlicht et al., 2014; Kool & Botvinick, 2014; Niyogi, Breton, et al., 2014; Niyogi, Shizgal, et al., 2014).

Results

Figure 2 plots the replay behavior of an agent pursuing a specified reward in a gridworld, using the EVB_C-driven replay algorithm described above, for different values of τ (all details of simulations are in the Method section of the Appendix). Three phases of replay behavior can be seen in these plots. In the first, the agent does not replay because there is no knowledge of the reward structure in the first trial, and thus there is no value to replay. Instead, the agent is accumulating experience to build an internal model of the environment (specifically, it is discovering rewards and developing its successor representation). In the second phase, the agent replays extensively because it has a good internal model but has still not fully developed and refined its value function, and thus can still gain by adjusting its Q-values through replay as well as action. Due to the expanded scope of replay (i.e., the ability to replay *any* experience), the speed benefit of replay, as well the low value of action, EVB_C is positive in this regime. This is the phase we identify with fatigue. Finally, 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 et al., 2016). This third regime—in which the value function has converged, EVB_C is negative, and behavior is executed without further deliberation—can be thought of as a transition to fully model-free, automatic processing. According to our model, cognitive fatigue, and corresponding periods of rest, arise during the preceding controlled, deliberative phase.

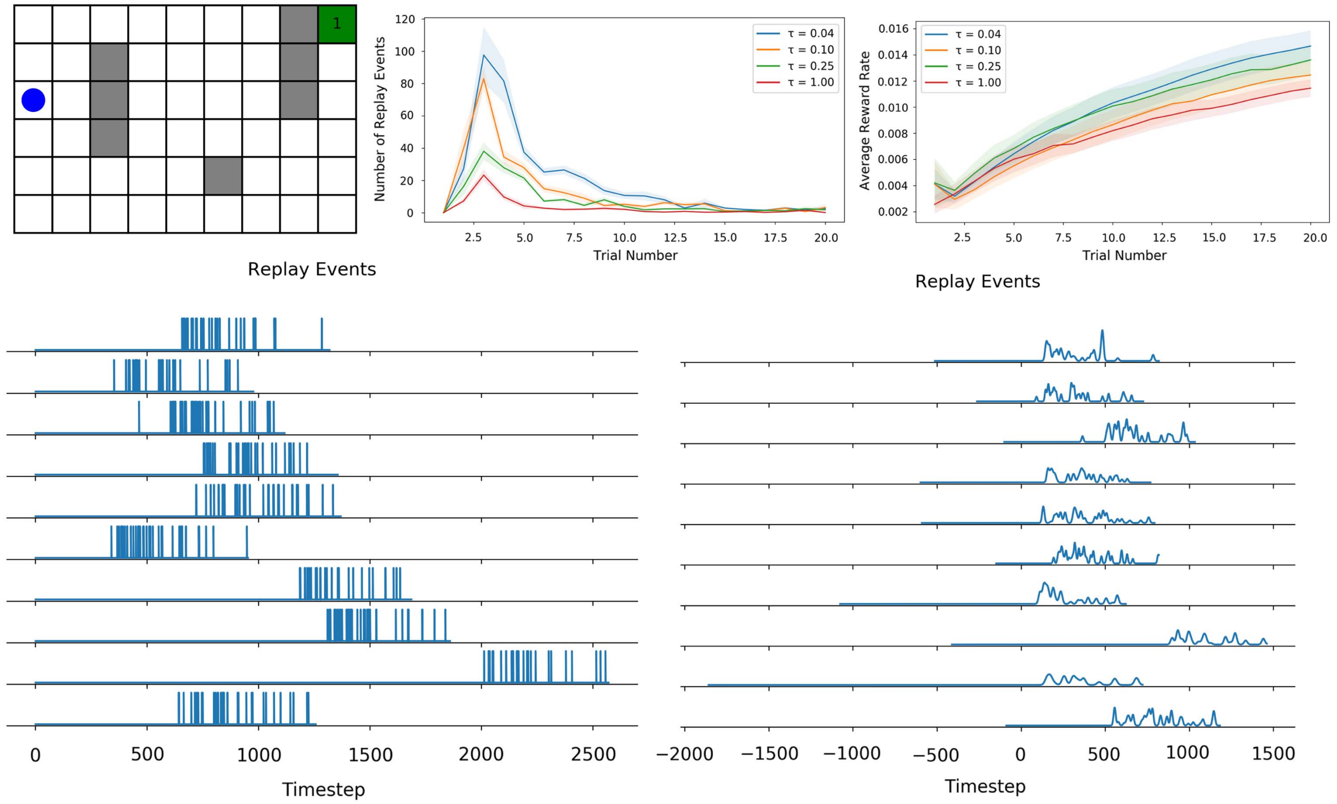
Our model explains the three canonical effects outlined earlier. The restorative power of rest demonstrated in Helton and Russell (2015) can be explained in a straightforward way in terms of replay: Rest provided the participants an opportunity for replay that facilitated learning and later performance (and also diminished the need for subsequent replay, which itself would compete with task performance). The 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

---

[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 60 times faster than physical action.

Figure 2

Simulation Results for Gridworld Agent



Note. (Top Left) Gridworld environment used for simulations. (Top Middle) Number of replay events over the course of multiple trials for different values of τ, the ratio of time it takes to replay versus act. (Top Right) Average reward rate for different values of τ. All error bars indicate ±1 Standard Error of the Mean (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). See the online article for the color version of this figure.

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 (Meyer & Kieras, 1997; Musslick et al., 2016; Navon & Gopher, 1979; Salvucci & Taatgen, 2008). 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 positing leisure as intrinsically valuable, we demonstrate how the temporal dynamics of its value as an opportunity for replay (mathematically derived in our model) leads it to be either greater than or less than the value of action at different points in time. Accordingly, a rational agent should arbitrate between periods of action and replay, based on which maximizes future reward.

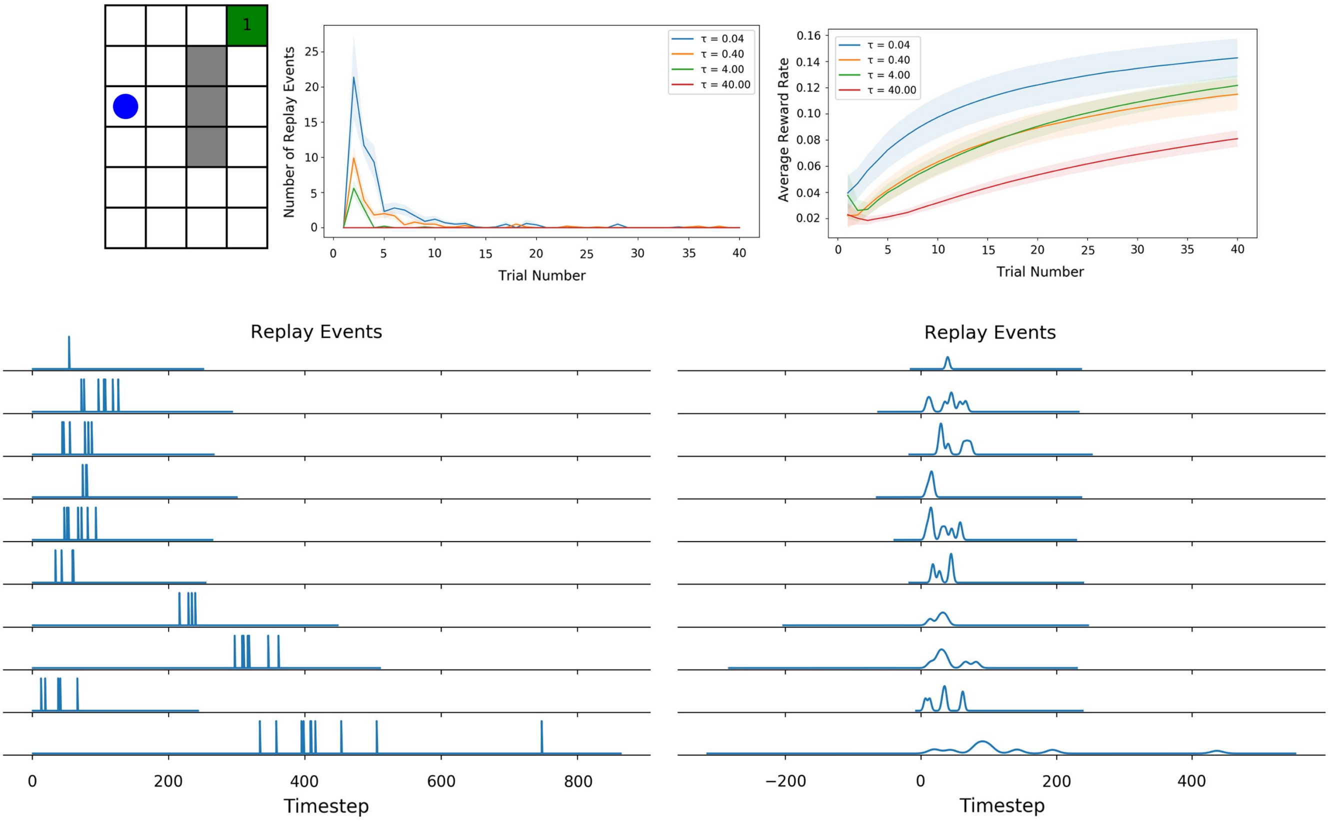
Finally, to model the relationship between task difficulty and fatigue, we evaluate our agent on an easier gridworld, shown in Figure 3. Since the agent takes less time to develop automaticity in this task (i.e., learn the relevant value function), the overall replay behavior (and hence fatigue) is reduced. 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

Figure 3

Simulation Results for Running Easy Gridworld Agent



Note. (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. Note that there is considerably less replay than for a more difficult task. (Top Right) Average reward rate for different values of τ. All error bars indicate ±1 Standard Error of the Mean (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). See the online article for the color version of this figure.

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 (Kumaran et al., 2016; McClelland et al., 1995; Schapiro et al., 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 et al. (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 multistep 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.

Intratrial Dynamics

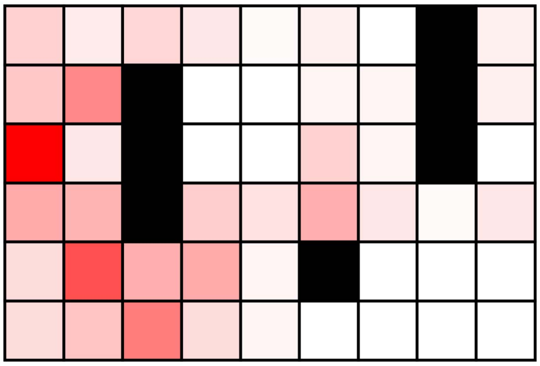
Fatigue studies generally consider the number of trials or the time on task as the causally relevant measure. The model proposed here suggests that learning is a mediating variable, and offers a more temporally fine-grained analysis, making quantitative predictions in terms of the states in a MDP. Consistent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

A Normative Lens to Understand Psychiatric Illnesses

Cognitive control and its attendant costs have been an important focus in the emerging field of computational psychiatry, which aims to give precise mathematical characterizations of mental illnesses in order to increase our understanding of these illnesses and move toward developing effective therapeutics (Huys et al., 2016; Montague et al., 2012; Wang & Krystal, 2014). Cognitive control is often considered to be disrupted in many psychiatric illnesses (Braver et al., 1999; Cohen & Servan-Schreiber, 1992), 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 remains 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 to develop a concrete and normative understanding of psychiatric illnesses. Consider posttraumatic 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 (Lieder et al., 2018; Mattar & Daw, 2018) and the current need/gain framework suggest that these should be sampled repeatedly and more often than nontraumatic 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 (Andrews & Thomson, 2009; Gagne et al., 2018; Kumaran et al., 2016).

Figure 4

Agent's Location When Replaying



Note. 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. See the online article for the color version of this figure.

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 versus a case of a more general function (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 et al., 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 et al., 2016; McCormick et al., 2020; Shin et al., 2019), 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.

Physical Fatigue

A natural extension of our work is to bring this framework into the domain of physical effort and fatigue. Admittedly, the semantic similarities between cognitive and physical fatigue do not necessarily imply a mechanistic similarity (in fact, some have argued that this perceived mechanistic relationship between physical and mental fatigue has been a distraction; Bartley & Chute, 1947; Hockey, 2011). There are clear physiological components to physical fatigue, which are beyond the scope of the current theory. However, there still may be mental and/or motivational components (Marcora & Staiano, 2010). Whether the effects of these factors are analogous to the role of mental simulation in cognitive tasks may be an exciting direction for future research.

Part 2: Boredom

The 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 overt 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 (Bench & Lench, 2013). Here, it seems that overt disengagement reflects disengagement from the task altogether, rather than a switch to a *covert* form of engagement in the service of improving the future performance of the current task.

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). The recently developed Meaning and Attentional Components (MAC) model (Westgate & Wilson, 2018) suggests that state boredom (Eastwood et al., 2012) is affected by two dissociable components: “Meaning” and “attention.” The “meaning” component corresponds to the alignment of the task with the agent’s goals (Van Tilburg & Igou, 2012), while the “attention” component corresponds to an alignment of the agent’s mental resources with the demands of the task (Eastwood et al., 2012; Hitchcock et al., 1999; London et al., 1972; Markey et al., 2014; Raffaelli et al., 2018; Wickens, 1991; Wickens, 2002).

Our aim in this Part is twofold. First, we develop a functional understanding of boredom by casting the insights of the MAC model in a utility-maximizing framework. Second, we use this formulation to provide insight on the second important issue addressed in this article: The *temporal dynamics* of boredom, that is, *why* boredom seems to increase during easy and/or repetitive tasks. To do so, we cast agents in an explore–exploit paradigm, and consider the idea that increases in boredom index the increasing 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 relative *value* of other tasks, it is perceived as a *cost* disfavoring status quo action; that is, it indexes the opportunity cost of foregoing exploration by a continued engagement in the current task. As we discuss below, the value of information (VOI) 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 three 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,” and the conjunction of meaning and attention), which is often at an intermediate level of difficulty; (2) if agents are bored, they will seek other tasks to perform; and, finally, (3) boredom increases over time while doing easy and/or repetitive tasks.

1. Optimal Participant-Task Fit

Functional theories consider boredom to arise when the current task is suboptimal for the agent, thus acting as a signal to disengage (Bench & Lench, 2013; Kurzban et al., 2013). Below, we provide a normative interpretation of the MAC model of boredom (Westgate & Wilson, 2018), suggesting that it elucidates situations in which the current task utility is low.

The first component, “meaning,” or the relevance of the task for the agent’s goals, directly corresponds to the notion of utility (reward, value) at the heart of reinforcement learning models. Tasks that align with agent’s goals have high utility, whereas tasks that do not have low utility. Low utility tasks such as copying references (Van Tilburg & Igou, 2012), counting words (Geana et al., 2016a), and passive number viewing (Milyavskaya et al., 2019) are often employed as boredom inductions in the literature. Manipulations increasing the “meaning” of a task, for example by incentivizing performance with charitable donations, reduce boredom even though the task remains the same (Westgate & Wilson, 2018).

The “attention” component corresponds to situations in which there is a mismatch between participant and task: Boredom occurs when demands are too high (“overstimulation”) or too low (“understimulation”). Tasks that are too difficult have low utility because the probability of success is low (Wickens, 1991, 2002), and thus participants feel bored when required to do a task they cannot do (Fisher, 1987, 1993; Hitchcock et al., 1999; Tanaka & Murayama, 2014). For example, Damrad-Frye and Laird (1989) distracted participants performing a comprehension task with extraneous noise and found that those in the distraction condition felt more bored than those in the no distraction condition.

We propose two reasons why “understimulation” leads to low utility. The first is related to opportunity costs: If an agent is doing an easy task, they can likely perform an additional task, which will have a greater combined utility than doing the sole easy task. Survey results have demonstrated that, to mitigate levels of boredom, many workers perform auxiliary tasks such as reading novels or writing letters (Fisher, 1987). Second, many “understimulating” tasks overlap with those considered to have low “meaning” (e.g., copying references), and thus also have low utility as noted above.

The proposed mapping between “meaning” and utility does not directly address one important question: What are the agent’s goals? That is, while it is generally agreed that copying references, counting words, and passive number viewing are not highly valued tasks, it is not explicitly clear *why* it is the case that an agent’s goals do not align with these tasks. More generally, decision-theoretic and reinforcement learning models often view utility as subjective, idiosyncratic to the agent, and do not offer a first-principle account of its source.

A reinforcement learning analysis can offer additional insight, relevant to boredom, as to how other, more distal aspects of a task may contribute to the motivation the agent has to perform the task. In particular, several recent lines of work have started to address this question, by using the VOI framework (Behrens et al., 2007; Bromberg-Martin & Hikosaka, 2009; Wilson et al., 2019) to suggest that the opportunity for learning is valuable. Agents not only value immediate reward, but also future (discounted) reward, and the VOI quantifies how gaining information, by improving future decisions, can increase expected future rewards when performing a task. Understimulating and/or low "meaning" tasks can often be considered to have a low VOI because there is little to no opportunity for learning available, and, in turn, low VOI for using any such knowledge to attain goals (utility) in the future.

A recent set of experiments by Geana et al. (2016a) directly evaluated this claim. 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. (2021).

2. *Switching to Other Tasks*

If the functional role of boredom is to signal the value of disengagement, we should see examples of boredom leading to task switching and general exploration. The second experiment of Geana et al. (2016a) sought to test this directly by allowing participants to switch voluntarily among the tasks used in their first experiment. They reasoned that if boredom is sensitive to the VOI, 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 current reward. Switching behavior can also be seen in human work environments, in which boredom has been found to lead to a higher labor turnover (Geiwitz, 1966; Kishida, 1973; Wild & Hill, 1970).

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 upfront, 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.

3. *The Temporal Dynamics of Boredom*

Boredom seems to increase over time when performing easy and/or repetitive tasks. Participants in the previously discussed Geana et al. (2016a) study increased their self-report levels of boredom as they engaged in the same task over a number of trials, and a similar effect was measured in Haager et al. (2018). These results support the general idea that tasks should increase in difficulty over time in order to maintain user engagement (Lawrence, 1952; Wilson et al., 2019), an insight widely leveraged by video games and curriculum designers. Understanding these temporal dynamics will shed insight on arguably the most ubiquitous, everyday experiences of state boredom: We often choose to perform a task precisely because it is not boring, but we eventually become bored of it (and thus choose to switch).

Discussion

Boredom can thus be considered as a state in which the current task has suboptimal utility, with the utility function comprised of the defined reward Q as well as the VOI. A bored agent should then disengage with the present task in order to pursue (or search for) one with higher overall utility, $Q_{\text{VOI}} = Q + \text{VOI}$. To explain the temporal dynamics, we propose that the change in boredom signals the changing VOI. 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 (Oudeyer et al., 2007; Schmidhuber, 1991; Wilson et al., 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 VOI can be expressed generically in a form analogous to EVB in the M&D replay model described in Part 1, providing a formal, integrated framework for investigating potential relationships between boredom and fatigue, as follows:

$$\text{VOI}(s, a_k) = \text{Gain}(s, a_k) \times \text{Need}(s, a_k), \tag{9}$$

in which

$$\text{Gain}(s, a_k) = \mathbb{E}\left[\sum_{a \in A} [\pi_{\text{new}}(a|s) - \pi_{\text{old}}(a|s)] q(s, a)\right], \tag{10}$$

$$= \int_{r \sim R(s,a_k)} p(r) \left[\sum_{a \in A} [\pi_{\text{new}}(a|s) - \pi_{\text{old}}(a|s)] q_{\text{new}}(s,a) \right] dr, \tag{11}$$

and

$$\text{Need}(s, a_k) = \sum_{i=1}^{\infty} \gamma^i \delta_{s_{t+i}, s}, \tag{12}$$

$$= \gamma \sum_{s' \in S} P(s, a_k, s') M(s', s), \tag{13}$$

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. 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, as captured by *Need*.[5]

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 50% chance of increasing by one and a 50% 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 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 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 (*understimulating*), as in Geana et al.'s Certain reward condition. The same thing happens for a task that is too hard (*overstimulating*); for example, rewards are stochastic but teach you nothing, as in Geana et al.'s Random condition in which there was no information to be gained.

The foregoing simulations describe dynamics of boredom related to uncertainty with respect to its effects on the Gain term. Our formulation also suggests boredom can be affected by the Need term. This can be examined, for instance, by manipulating the horizon in multi-armed bandit settings. For example, Figure 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 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 VOI 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: Understimulating and overstimulation (tasks that are boring due to being too easy or too hard) both correspond to situations in which the task's Q_{VOI} is low. Furthermore, when the agent is bored, that is, 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 to which our formulation can potentially connect.

Adaptive Gain Theory

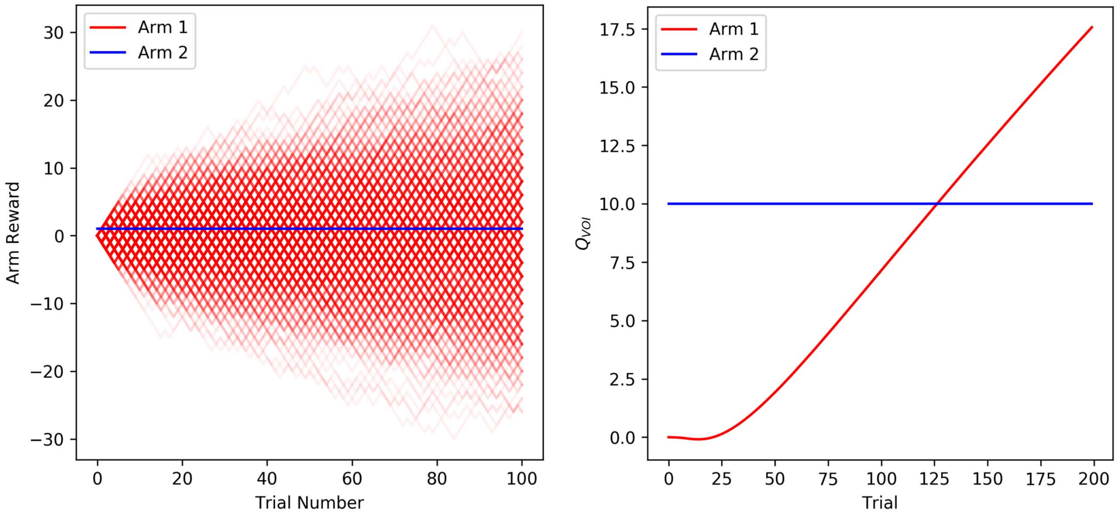
While our analysis has primarily been at Marr's (1982) computational level, a complete theory would account for effects at all levels of analysis. One line of work that has pursued both a mechanistic and normative account of arousal has focused on its association with norepinephrine (NE) function—a neuromodulator that is widely distributed throughout the brain. The Adaptive Gain Theory (AGT; Aston-Jones & Cohen, 2005) of NE function suggests that low/medium/high levels of arousal correspond to

---

[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, they do not consider the additional informational value of additional learning at these subsequent steps. This approximation was chosen to match the same simplification as used for EVB by M&D, as in the previous section, and is sufficient for our purposes.

Figure 5

Sample Bandit Task Illustrating the Increase in Q_{VOI} With Increased Uncertainty

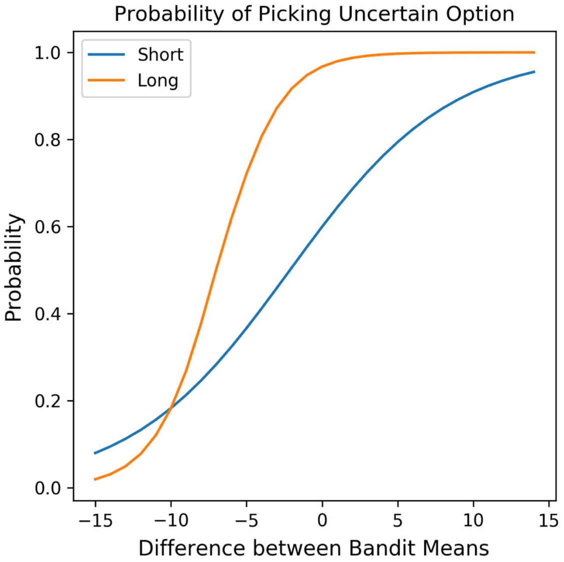


Note. (Left) Red indicates sample reward trajectories of the first arm, the reward for which starts at zero and then increases or decreases by one every iteration. The blue arm 2 indicates an arm with value always at 1. (Right) The change in Q_{VOI} over time if the agent keeps on picking the blue arm. 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. See the online article for the color version of this figure.

low/medium/high levels of sustained NE release. AGT argues that high tonic levels of NE can be seen as favoring disengagement from the current task set in favor of switching to some other one. Recently, Kane et al. (2017) presented causal evidence supporting this hypothesized relationship. In this work, the authors found that increasing tonic NE levels in locus coeruleus (LC; the brainstem nucleus that is the source of NE) led rodents to explore more in a patch-foraging task. NE levels may thus provide a mechanism by which the brain implements the VOI 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 (Dabney et al., 2020; Schultz, 1998).

Figure 6

Probability of Uncertain Option Increases in Long Horizon



Note. We reproduce the effect found in 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. See the online article for the color version of this figure.

Curiosity

We have focused on boredom as a negative reflection of the VOI, which drives choices away from uninformative options. It is clearly also the case that there are also affective states associated in a positive way with the informational value of exploration, and may play a complementary role in directing choices toward informative options. 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. Thus, one way to dissociate boredom from curiosity is that the former prompts disengagement from uninformative tasks, while the latter prompts engagement with informative tasks.

The emergence of these formal models will hopefully permit future work aimed at determining the extent to which curiosity versus 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 the features of VOI to which they are sensitive? Finally, we have argued (and Geana et al.'s (2016a) results support) that boredom is a *differential* signal of Q_{VOI}, in the sense that it is increased not just by (low) Q_{VOI} for the current task but also by (high) Q_{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 VOI. 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 nonstationary, 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 $Q_{\text{VOI}} = Q + \text{VOI}$ 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 (Moore & Atkeson, 1993; Peng & Williams, 1993):

$$\begin{aligned}\text{Gain}(s_t, a_t)\\ = \rho|R(s_t, a_t) + \gamma \max_a Q_{\text{VOI}}(s_{t+1}, a) - Q_{\text{VOI}}(s_t, a_t)|, \quad (14)\end{aligned}$$

which is the 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.[6] We also include an optional degree of freedom $\rho \leq 1$ to scale the heuristic, though we set $\rho = 1$ for our simulations.

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

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

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 VOI 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 nonvisited 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.

Discussion

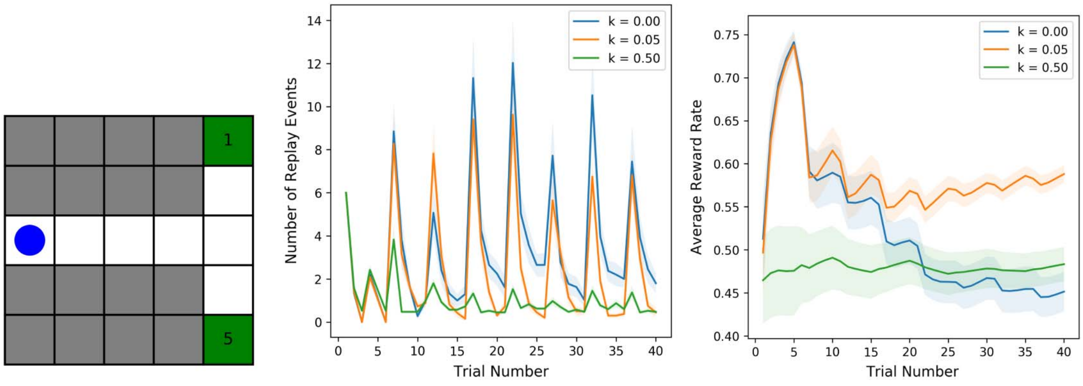
It has sometimes been questioned whether performance decrements following long time-on-task are a result of fatigue or boredom (Mackworth, 1968; Pattyn et al., 2008). Meanwhile, other work has not discriminated between these states, implicitly considering both as reflecting motivation and/or opportunity 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

---

[6] 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



Note. 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 Standard Error of the Mean (SEM). See the online article for the color version of this figure.

due to uncertainty and VOI now driving replay as well as exploratory action. This remains perpetually high because we have assumed (as in the boredom modeling) that the value of terminal states may change if they are long unvisited. Since these relate to both EVB and VOI, it is unclear within our framework which phenomenological state to associate with each.

It is also possible that such interactions could be associated with other phenomenological states. Above, we discussed curiosity as one possibility. Another is "mind-wandering," the act of spontaneously generating thoughts, that has become a focus of empirical study given its prevalence in daily life (Christoff et al., 2016; Danckert & Merrifield, 2018; Fox & Christoff, 2014; Fox et al., 2013; Smallwood & Schooler, 2006; Smallwood & Schooler, 2015). Recent work has suggested that mind-wandering can be goal-directed, and can facilitate creativity (Agnoli et al., 2018; Fox & Christoff, 2018; Williams et al., 2018; Zedelius & Schooler, 2015). 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 such tasks that can be performed at once is limited. This implies 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 task 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 more boring. Boredom is minimized in "flow" states, in which there is both a high exploitative and high exploratory value in the current task.

While this is a subtle distinction, it is a potentially important one, that reflects a primary goal of our effort: To tie the constructs of boredom and fatigue to *distinct* computational mechanisms, in a way that provides a formally precise and empirically testable explanation for two related, but distinguishable ways in which task engagement can decrease over time. This formulation is inspired by and seeks to explain the subjective phenomenology associated with the terms boredom and fatigue, which we assume reflect computational signals generated by the proposed mechanisms. However, testing this relationship is challenging, given that self-report—the primary means of measuring the phenomenology—may not reliably reflect the underlying mechanisms.

Accordingly, while we hope our theory can explain a sufficient number of findings concerning boredom and fatigue to warrant consideration, it is reasonable to expect that it will not account for all of them. For example, Milyavskaya et al. (2019) surprisingly found that their boredom induction increased participants' self-report ratings of fatigue more than their effort manipulation did.

One possible explanation for this result is that that the self-report scale used "fatigued" and "energized" as the two endpoints, but participants might not consider those to be opposite ends of the same phenomenon. Another tentative explanation is that the bored participants were mind-wandering (e.g., Danckert & Merrifield, 2018) and that the act of mind-wandering increases fatigue. Future work will be required to resolve whether deviations between theory and measures of the subjective report in the literature reflect a failure of the theory, or imprecisions in self-report that it may help overcome. Toward that end, one important approach will be to evaluate other forms of measurement that may correlate with boredom and/or fatigue (e.g., pupil diameter or neural signals) and that may provide more proximal markers of the internal computational signals proposed by the theory (Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011).

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 multihour *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. While both conditions are clearly both boring and fatiguing, our computational framework specifies why their relative degree should differ between conditions. When opportunity costs are increasing at a greater rate during a more control-dependent task, fatigue is responsible. If the increase in impulsive choices had instead measured boredom, we would have expected the opposite: Those in the easier task condition would change their impulsivity more than those in the harder task. We predict that relatively greater boredom in the easier task condition might indeed be captured using a different dependent measure, such as pupillometry or exploratory choices on a subsequent bandit task. Note also that, under our theory, time discounting would be expected to reflect fatigue only to the extent that patient behavior on delay discounting requires neural operations that compete with replay of the *N*-back task. While we are not aware of evidence that directly tests this claim, it is consistent with suggestions that intertemporal choice overlaps mechanistically with future-oriented deliberation (Hunter et al., 2018; Peters & Büchel, 2010), for example, because patient choices are promoted by adequate mental simulation of their salutary consequences.

Finally, 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 VOI 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 and 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 (Moore & Atkeson, 1993; Peng & Williams, 1993; Schaul et al., 2015) and neuroscience (Momennejad et al., 2018) literature, but it remains an open question as to how the brain computes these values. Similarly, the VOI 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 (Wilson et al., 2019), a heuristic that 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 et al., 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 et al., 2009; Jadhav et al., 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 opportunity costs: Mental simulation (replay) and exploration, which 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 (Howes et al., 2009; Lieder & Griffiths, 2020; Shenhav et al., 2013; Simon, 1972), 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.

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. https://doi.org/10.1080/10400419.2018.1411423

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem situations. *Psychological review*, *94*(2), 192–210. https://doi.org/10.1037/0033-295X.94.2.192

Andrews, P. W., & Thomson, J. A. (2009). The bright side of being blue: depression as an adaptation for analyzing complex problems. *Psychological review*, *116*(3), 620–654. https://doi.org/10.1037/a0016242

Antony, J. W., Hartshorne, T. H., Pomeroy, K., Gureckis, T. M., Hasson, U., McDougle, S. D., & Norman, K. A. (2021). Behavioral, physiological, and neural signatures of surprise during naturalistic sports viewing. *Neuron*, *109*(2), 377–390. https://doi.org/10.1016/j.neuron.2020.10.029

Arai, T. (1912). *Mental fatigue*. Teachers College, Columbia University.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-Time analysis of the multiarmed bandit problem. *Machine learning*, *47*(2–3), 235–256. https://doi.org/10.1023/A:1013689704352

Bartley, S. H., & Chute, E. (1947). *Fatigue and impairment in man*. McGraw-Hill Book Company.

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), 1252–1265. https://doi.org/10.1037//0022-3514.74.5.1252

Baumeister, R. F., & Vohs, K. D. (2007). Self-Regulation, ego depletion, and motivation. *Social and personality psychology compass*, *1*(1), 115–128. https://doi.org/10.1111/j.1751-9004.2007.00001.x

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Bench, S. W., & Lench, H. C. (2013). On the function of boredom. *Behavioral sciences*, *3*(3), 459–472. https://doi.org/10.3390/bs3030459

Bergum, B. O., & Lehr, D. J. (1962). Vigilance performance as a function of interpolated rest. *Journal of Applied Psychology*, *46*(6), 425–427. https://doi.org/10.1037/h0043505

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. https://doi.org/10.1073/pnas.1520527113

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. https://doi.org/10.1080/00220671.1981.10885317

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. https://doi.org/10.1016/s0006-3223(99)00116-x

Bromberg-Martin, E. S., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*(1), 119–126. https://doi.org/10.1016/j.neuron.2009.06.009

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), 147–153. https://doi.org/10.1038/nn.2732

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), 796–815. https://doi.org/10.1037/xge0000083

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), 718–731. https://doi.org/10.1038/nrn.2016.113

Cohen, G., & Burke, D. M. (1993). Memory for proper names: A review. *Memory*, *1*(4), 249–263. https://doi.org/10.1080/09658219308258237

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), 332–361. https://doi.org/10.1037/0033-295x.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. https://doi.org/10.1098/rstb.2007.2098

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychological review*, *99*(1), 45–77. https://doi.org/10.1037/0033-295x.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. https://doi.org/10.1523/JNEUROSCI.22-24-10976.2002

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

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*, *577*(7792), 671–675. https://doi.org/10.1038/s41586-019-1924-6

Damrad-Frye, R., & Laird, J. D. (1989). The experience of boredom: The role of the self-perception of attention. *Journal of Personality and Social Psychology*, *57*(2), 315–320. https://doi.org/10.1037/0022-3514.57.2.315

Danckert, J., & Merrifield, C. (2018). Boredom, sustained attention and the default mode network. *Experimental brain research*, *236*(9), 2507–2518. https://doi.org/10.1007/s00221-016-4617-5

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*, Article 321497. https://doi.org/10.1101/321497

Davidson, T. J., Kloosterman, F., & Wilson, M. A. (2009). Hippocampal replay of extended experience. *Neuron*, *63*(4), 497–507. https://doi.org/10.1016/j.neuron.2009.07.027

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

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. https://doi.org/10.1038/nature04766

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624. https://doi.org/10.1162/neco.1993.5.4.613

Diba, K., & Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature neuroscience*, *10*(10), 1241–1242. https://doi.org/10.1038/nn1961

Dodge, R. (1917). The laws of relative fatigue. *Psychological Review*, *24*(2), 89–113. https://doi.org/10.1037/h0075549

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), 795–805. https://doi.org/10.1037/0021-9010.84.5.795

Dora, J., van Hooff, M., Geurts, S., Kompier, M., & Bijleveld, E. (2019). The effect of opportunity costs on mental fatigue in labor/leisure tradeoffs. *PsyArXiv*. https://doi.org/10.31234/osf.io/3765s

Dubey, R., & Griffiths, T. (2019). Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review*, *127*(3), 455–476. https://doi.org/10.1037/rev0000175

Eastwood, J. D., Frischen, A., Fenske, M. J., & Smilek, D. (2012). The unengaged mind: Defining boredom in terms of attention. *Perspectives on Psychological Science*, *7*(5), 482–495. https://doi.org/10.1177/1745691612456044

Eldar, E., Lièvre, G., Dayan, P., & Dolan, R. J. (2020). The roles of online and offline replay in planning. *BioRxiv*. https://doi.org/10.7554/eLife.56911

Fisher, C. D. (1987). *Boredom: Construct, causes and consequences* (Tech. Rep.). Texas A&M University College Station Department of Management.

Fisher, C. D. (1993). Boredom at work: A neglected concept. *Human Relations*, *46*(3), 395–417. https://doi.org/10.1177/001872679304600305

Foster, D. J., & Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, *440*(7084), 680–683. https://doi.org/10.1038/nature04587

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*, Article 412. https://doi.org/10.3389/fnhum.2013.00412

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. https://doi.org/10.1016/j.cobeha.2018.03.013

Geana, A., Wilson, R., Daw, N. D., & Cohen, J. D. (2016a). *Boredom, information-seeking and exploration* [Conference session]. In Proceedings of the 38th Annual Conference of the Cognitive Science Society, Philadelphia, PA, United States.

Geana, A., Wilson, R., Daw, N. D., & Cohen, J. D. (2016b). *Information-seeking, learning and the marginal value theorem: A normative approach to adaptive* [Conference session]. In Proceedings of the 38th Annual Conference of the Cognitive Science Society, Philadelphia, PA, United States.

Geiwitz, P. J. (1966). Structure of boredom. *Journal of personality and social psychology*, *3*(5), 592–600. https://doi.org/10.1037/h0023202

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42. https://doi.org/10.1016/j.cognition.2017.12.014

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), 182–194. https://doi.org/10.1037/a0030844

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. https://doi.org/10.1162/NECO_a_00282

Gibson, E. L. (2007). Carbohydrates and mental function: Feeding or impeding the brain? *Nutrition Bulletin*, *32*, 71–83. https://doi.org/10.1111/j.1467-3010.2007.00606.x

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective & Behavioral Neuroscience*, *10*(2), 252–269. https://doi.org/10.3758/CABN.10.2.252

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), 1222–1223. https://doi.org/10.1038/nn.2384

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(2), 148–164. https://doi.org/10.1214/aoap/1177005588

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. https://doi.org/10.1016/j.neuron.2010.01.034

Haager, J. S., Kuhbandner, C., & Pekrun, R. (2018). To be bored or not to be bored—how task-related boredom influences creative performance. *The Journal of Creative Behavior*, *52*(4), 297–304. https://doi.org/10.1002/jocb.154

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., . . . Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1002/jocb.154

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), 188–199. https://doi.org/10.1037/1076-898X.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. https://doi.org/10.1016/j.cognition.2014.10.001

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. https://doi.org/10.1177/0018720816683509

Hitchcock, E. M., Dember, W. N., Warm, J. S., Moroney, B. W., & See, J. E. (1999). Effects of cueing and knowledge of results on workload and boredom in sustained attention. *Human factors*, *41*(3), 365–372. https://doi.org/10.1518/001872099779610987

Hockey, G. R. J. (2011). A motivational control theory of cognitive fatigue. In P. L. Ackerman (Ed.), *Decade of behavior/science conference. Cognitive fatigue: Multidisciplinary perspectives on current research and future applications* (pp. 167–187). American Psychological Association.

Hockey, R. (2013). *The psychology of fatigue: Work, effort and control*. Cambridge University Press.

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), 717–751. https://doi.org/10.1037/a0017187

Hunter, L. E., Bornstein, A. M., & Hartley, C. A. (2018). A common deliberative process underlies model-based planning and patient intertemporal choice. *bioRxiv*. Article 499707. https://doi.org/10.1101/499707

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. https://doi.org/10.1037/h0093556

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, *19*(3), 404–413. https://doi.org/10.1038/nn.4238

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. https://doi.org/10.1016/j.tics.2013.12.009

Izawa, C. (1971). Massed and spaced practice in paired-associate learning: List versus item distributions. *Journal of Experimental Psychology*, *89*(1), 10–21. https://doi.org/10.1037/h0031177

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. https://doi.org/10.1126/science.1217230

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. https://doi.org/10.1126/science.1217230

Jepma, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration-exploitation trade-off: Evidence for the adaptive gain theory. *Journal of Cognitive Neuroscience*, *23*(7), 1587–1596. https://doi.org/10.1162/jocn.2010.21548

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, *4*, 237–285. https://doi.org/10.1613/jair.301

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. https://doi.org/10.3758/s13415-017-0531-y

Karlsson, M. P., & Frank, L. M. (2009). Awake replay of remote experiences in the hippocampus. *Nature neuroscience*, *12*(7), 913–918. https://doi.org/10.1038/nn.2344

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

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. https://doi.org/10.1073/pnas.1609094113

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, *55*(4), 352–358. https://doi.org/10.1037/h0043688

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*, *143*(1), 131–141. https://doi.org/10.1037/a0031048

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLoS computational biology*, *12*(8), Article e1005090. https://doi.org/10.1371/journal.pcbi.1005090

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-Benefit arbitration between multiple reinforcement-learning systems. *Psychological science*, *28*(9), 1321–1333. https://doi.org/10.1177/0956797617708288

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. https://doi.org/10.1016/j.tics.2016.05.004

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. https://doi.org/10.1017/S0140525X12003196

Lawrence, D. H. (1952). The transfer of a discrimination along a continuum. *Journal of Comparative and Physiological Psychology*, *45*(6), 511–516. https://doi.org/10.1037/h0057135

Lee, A. K., & Wilson, M. A. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, *36*(6), 1183–1194. https://doi.org/10.1016/s0896-6273(02)01096-6

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. https://doi.org/10.1016/j.neuron.2013.11.028

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*. https://doi.org/10.1017/S0140525X1900061X

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), 1–32. https://doi.org/10.1037/rev0000074

Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. (2019). Human replay spontaneously reorganizes experience. *Cell*, *178*(3), 640–652. https://doi.org/10.1016/j.cell.2019.06.012

Liu, Y., Mattar, M., Behrens, T., Daw, N., & Dolan, R. J. (2020). Experience replay supports non-local learning. *Science (New York, N.Y.)*, *372*(6544), Article eabf1357. https://doi.org/10.1126/science.abf1357

London, H., Schubert, D. S., & Washburn, D. (1972). Increase of autonomic arousal by boredom. *Journal of Abnormal Psychology*, *80*(1), 29–36. https://doi.org/10.1037/h0033311

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. https://doi.org/10.1016/j.cogbrainres.2005.01.018

Louie, K., & Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, *29*(1), 145–156. https://doi.org/10.1016/S0896-6273(01)00186-6

Mackworth, J. F. (1968). Vigilance, arousal, and habituation. *Psychological review*, *75*(4), 308–322. https://doi.org/10.1037/h0025896

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, *1*(1), 6–21. https://doi.org/10.1080/17470214808416738

Marcora, S. M., & Staiano, W. (2010). The limit to exercise tolerance in humans: mind over muscle? *European journal of applied physiology*, *109*(4), 763–770. https://doi.org/10.1007/s00421-010-1418-6

Markey, A., Chin, A., Vanepps, E. M., & Loewenstein, G. (2014). Identifying a reliable boredom induction. *Perceptual and motor skills*, *119*(1), 237–253. https://doi.org/10.2466/27.PMS.119c18z6

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

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience*, *21*(11), 1609–1617. https://doi.org/10.1038/s41593-018-0232-z

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), 419–457. https://doi.org/10.1037/0033-295X.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. *Cerebral Cortex. 30*(11), 5972–5987. https://doi.org/10.1093/cercor/bhaa172

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Daniel, H., Klaus, F., & Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191–215. https://doi.org/10.1037/dec0000033

Messier, C. (2004). Glucose improvement of memory: a review. *European journal of pharmacology*, *490*(1–3), 33–57. https://doi.org/10.1016/j.ejphar.2004.02.043

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), 978–984. https://doi.org/10.1037/xlm0000216

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), 3–65. https://doi.org/10.1037/0033-295x.104.1.3

Milyavskaya, M., Inzlicht, M., Johnson, T., & Larson, M. J. (2019). Reward sensitivity following boredom and cognitive effort: A high-powered neurophysiological investigation. *Neuropsychologia*, *123*, 159–168. https://doi.org/10.1016/j.neuropsychologia.2018.03.033

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-Level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

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. https://doi.org/10.1177/0956797612439069

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *Elife*, *7*. Article e32548. https://doi.org/10.7554/eLife.32548

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. https://doi.org/10.1523/JNEUROSCI.16-05-01936.1996

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, *16*(1), 72–80. https://doi.org/10.1016/j.tics.2011.11.018

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

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, *31*, 69–89. https://doi.org/10.1146/annurev.neuro.31.061307.090723

Müller, T., & Apps, M. A. (2019). Motivational fatigue: A neurocognitive framework for the impact of effortful exertion on subsequent motivation. *Neuropsychologia*, *123*, 141–151. https://doi.org/10.1016/j.neuropsychologia.2018.04.030

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* [Conference session]. In Proceedings of the 38th Annual Conference of the Cognitive Science Society, Philadelphia, PA, United States.

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. https://doi.org/10.1523/JNEUROSCI.19-21-09497.1999

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological review*, *86*(3), 214–255. https://doi.org/10.1037/0033-295X.86.3.214

Newberg, A. B., Wang, J., Rao, H., Swanson, R. L., Wintering, N., Karp, J. S., Alavi, A., Greenberg, J. H., & Detre, J. A. (2005). Concurrent cbf and cmrglc changes during human brain activation by combined fmri–pet scanning. *Neuroimage*, *28*(2), 500–506. https://doi.org/10.1016/j.neuroimage.2005.06.040

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), Article 969. https://doi.org/10.1098/rsif.2013.0969

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), Article e1003894. https://doi.org/10.1371/journal.pcbi.1003894

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain research*, *34*, 171–175. https://doi.org/10.1016/0006-8993(71)90358-1

Ólafsdóttir, H. F., Bush, D., & Barry, C. (2018). The role of hippocampal replay in memory and planning. *Current Biology*, *28*(1), R37–R50. https://doi.org/10.1016/j.cub.2017.10.073

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. https://doi.org/10.1177/0956797612463080

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. https://doi.org/10.1073/pnas.1312011110

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. https://doi.org/10.1109/TEVC.2006.890271

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. https://doi.org/10.1016/j.physbeh.2007.09.016

Peng, J., & Williams, R. J. (1993). Efficient learning and planning within the dyna framework. *Adaptive Behavior*, *1*(4), 437–454. https://doi.org/10.1177/105971239300100403

Peters, J., & Büchel, C. (2010). Episodic future thinking reduces reward delay discounting through an enhancement of prefrontal-mediotemporal interactions. *Neuron*, *66*(1), 138–148. https://doi.org/10.1016/j.neuron.2010.03.026

Pfeiffer, B. E., & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, *497*(7447), 74–79. https://doi.org/10.1038/nature12112

Posner, M., & Snyder, C. R. (1975). Facilitation and inhibition in the processing of signals. In P. M. Rabbitt, & S. Dornic, (Eds.), *Attention and Performance* (Vol. 5, pp. 669–682) Academic Press.

Raffaelli, Q., Mills, C., & Christoff, K. (2018). The knowns and unknowns of boredom: A review of the literature. *Experimental brain research*, *236*(9), 2451–2462. https://doi.org/10.1007/s00221-017-4922-7

Raichle, M. E., & Mintun, M. A. (2006). Brain work and brain imaging. *Annual Review of Neuroscience*, *29*, 449–476. https://doi.org/10.1146/annurev.neuro.29.051605.112819

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), Article e0182980. https://doi.org/10.1371/journal.pone.0182980

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*, *4*(1), 11–18.

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. https://doi.org/10.1007/s00221-014-3865-5

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological review*, *115*(1), 101–130. https://doi.org/10.1037/0033-295X.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), Article 3920. https://doi.org/10.1038/s41467-018-06213-1

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), Article 20160049. https://doi.org/10.1098/rstb.2016.0049

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), 551–566. https://doi.org/10.1037/a0029487

Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the international conference on simulation of adaptive behavior: From animals to animats* (pp. 222–227). MIT Press.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. detection, search, and attention. *Psychological review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Schuck, N. W., & Niv, Y. (2019). Sequential replay of nonspatial task states in the human hippocampus. *Science*, *364*(6447), Article eaaw5181. https://doi.org/10.1126/science.aaw5181

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, *80*(1), 1–27. https://doi.org/10.1152/jn.1998.80.1.1

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

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. https://doi.org/10.1073/pnas.1821028116

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), Article e1006827. https://doi.org/10.1371/journal.pcbi.1006827

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. https://doi.org/10.1016/j.neuron.2013.07.007

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), 127–190. https://doi.org/10.1037/0033-295X.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*, *104*(6), 1110–1125. https://doi.org/10.1016/j.neuron.2019.09.012

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, Pittsburgh, PA, United States* (pp. 833–840).

Smallwood, J., & Schooler, J. W. (2006). The restless mind. *Psychological bulletin*, *132*(6), 946–958. https://doi.org/10.1037/0033-2909.132.6.946

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual review of psychology*, *66*, 487–518. https://doi.org/10.1146/annurev-psych-010814-015331

Solway, A., & Botvinick, M. M. (2012). Goal-Directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychological review*, *119*(1), 120–154. https://doi.org/10.1037/a0026435

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

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*. 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. https://doi.org/10.1016/j.neuron.2010.01.001

Tanaka, A., & Murayama, K. (2014). Within-Person analyses of situational interest and boredom: Interactions between task-specific perceptions and achievement goals. *Journal of Educational Psychology*, *106*(4), 1122–1134. https://doi.org/10.1037/a0036659

Thorndike, E. (1900). Mental fatigue. i. *Psychological Review*, *7*(5), 466–482. https://doi.org/10.1037/h0069440

Thorndike, E. L. (1912). The curve of work. *Psychological Review*, *19*(3), 165–194. https://doi.org/10.1037/h0073541

Tononi, G., & Cirelli, C. (2003). Sleep and synaptic homeostasis: A hypothesis. *Brain research bulletin*, *62*(2), 143–150. https://doi.org/10.1016/j.brainresbull.2003.09.004

Tononi, G., & Cirelli, C. (2006). Sleep function and synaptic homeostasis. *Sleep medicine reviews*, *10*(1), 49–62. https://doi.org/10.1016/j.smrv.2005.05.002

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. https://doi.org/10.1016/j.neuron.2013.12.025

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. https://doi.org/10.1016/j.neuron.2016.10.020

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. https://doi.org/10.1016/s0001-6918(02)00150-6

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*. Article 1. https://doi.org/10.3389/neuro.07.001.2009

Van Tilburg, W. A., & Igou, E. R. (2012). On boredom: Lack of challenge and meaning as distinct boredom experiences. *Motivation and Emotion*, *36*(2), 181–194. https://doi.org/10.1007/s11031-011-9234-9

Wamsley, E. J. (2019). Memory consolidation during waking rest. *Trends in cognitive sciences*, *23*(3), 171–173. https://doi.org/10.1016/j.tics.2018.12.007

Wang, X.-J., & Krystal, J. H. (2014). Computational psychiatry. *Neuron*, *84*(3), 638–654. https://doi.org/10.1016/j.neuron.2014.10.018

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human factors*, *50*(3), 433–441. https://doi.org/10.1518/001872008X312152

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3–4), 279–292. https://doi.org/10.1007/BF00992698

Westgate, E. C., & Wilson, T. D. (2018). Boring thoughts and bored minds: The mac model of boredom and cognitive engagement. *Psychological Review*, *125*(5), 689–713. https://doi.org/10.1037/rev0000097

Wickens, C. D. (1991). Processing resources and attention. In *Multiple-task performance* (pp. 3–34). CRC Press.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, *3*(2), 159–177. https://doi.org/10.1080/14639220210123806

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

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. https://doi.org/10.1016/j.jenvp.2018.08.005

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, *265*(5172), 676–679. https://doi.org/10.1126/science.8036517

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), 2074–2081. https://doi.org/10.1037/a0038199

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. *Nature Neuroscience*, *23*(8), 1025–1033. https://doi.org/10.1038/s41593-020-0649-z

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. https://doi.org/10.1002/cne.920180503

Zedelius, C. M., & Schooler, J. W. (2015). Mind wandering "ahas" versus mindful reasoning: Alternative routes to creative solutions. *Frontiers in Psychology*, *6*. Article 834. https://doi.org/10.3389/fpsyg.2015.00834

(*Appendix follows*)

Appendix

Method

Part 1: Cognitive Fatigue

$$\begin{aligned}
&\mathrm{EVB}_C(s, s_k, a_k)\\
&= \mathop{\mathbb{E}}_{\pi_{\text{new}}}\left[\sum_{i=\tau}^{\infty}\gamma^i R_{t+i}|S_t = s\right] - \mathop{\mathbb{E}}_{\pi_{\text{old}}}\left[\sum_{i=0}^{\infty}\gamma^i R_{t+i}|S_t = s\right], && (16)\\
&= \gamma^{\tau} v_{\pi_{\text{new}}}(s) - v_{\pi_{\text{old}}}(s), && (17)\\
&= \gamma^{\tau} v_{\pi_{\text{new}}}(s) - \gamma^{\tau} v_{\pi_{\text{old}}}(s) + \gamma^{\tau} v_{\pi_{\text{old}}}(s) - v_{\pi_{\text{old}}}(s), && (18)\\
&= \gamma^{\tau}(v_{\pi_{\text{new}}}(s) - v_{\pi_{\text{old}}}(s)) + \gamma^{\tau} v_{\pi_{\text{old}}}(s) - v_{\pi_{\text{old}}}(s), && (19)\\
&= \gamma^{\tau}\mathrm{EVB}(s, s_k, a_k) + \gamma^{\tau} v_{\pi_{\text{old}}}(s) - v_{\pi_{\text{old}}}(s), && (20)\\
&= \gamma^{\tau}\mathrm{EVB}(s, s_k, a_k) - (1 - \gamma^{\tau}) v_{\pi_{\text{old}}}(s), && (21)\\
&= \gamma^{\tau}\mathrm{EVB}(s, s_k, a_k) - (1 - \gamma^{\tau})\sum_{a}\pi_{\text{old}}(a|s)q_{\pi_{\text{old}}}(s, a). && (22)
\end{aligned}$$

A previous version of the manuscript incorrectly decomposed EVB_C into Gain × Need in which,

$$\begin{aligned}
&\mathrm{Gain}(s_k, a_k)\\
&= \gamma^{\tau}\sum_{a\in A}Q_{\pi_{\text{new}}}(s_k, a)\pi_{\text{new}}(a|s_k) - \sum_{a\in A}Q_{\pi_{\text{new}}}(s_k, a)\pi_{\text{old}}(a|s_k), && (23)
\end{aligned}$$

This mistake has been corrected in the present manuscript.

The original gridworld comprised of a 9×6 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:

$$\pi(a|s) = \frac{e^{\beta Q(s,a)}}{\sum_{d} e^{\beta Q(s,d)}}, \tag{24}$$

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 Equation 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 4×6 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 3 plot the replay behavior of the 10 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 50% chance of changing by plus one and 50% chance of changing by minus one for 20 iterations. For the short horizon, a horizon of 2 was used, while a horizon of 10 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 5×5 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 Equation 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 40 episodes were run for each value of k.

Received September 8, 2020
Revision received April 8, 2021
Accepted May 9, 2021 ■