LETTER **Communicated by Michael Mozer**

Estimating Scale-Invariant Future in Continuous Time

Zoran Tiganj
zoran.tiganj@gmail.com
Center for Memory and Brain, Department of Psychological and Brain Sciences, Boston, MA 02215, U.S.A.

Samuel J. Gershman
gershman@fas.harvard.edu
Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138, U.S.A.

Per B. Sederberg
pbs5u@virginia.edu
Department of Psychology, University of Virginia, Charlottesville, VA, 22904, U.S.A.

Marc W. Howard
marc777@bu.edu
Center for Memory and Brain, Department of Psychological and Brain Sciences, Boston, MA 02215, U.S.A.

Natural learners must compute an estimate of future outcomes that follow from a stimulus in continuous time. Widely used reinforcement learning algorithms discretize continuous time and estimate either transition functions from one step to the next (model-based algorithms) or a scalar value of exponentially discounted future reward using the Bellman equation (model-free algorithms). An important drawback of model-based algorithms is that computational cost grows linearly with the amount of time to be simulated. An important drawback of model-free algorithms is the need to select a timescale required for exponential discounting. We present a computational mechanism, developed based on work in psychology and neuroscience, for computing a scale-invariant timeline of future outcomes. This mechanism efficiently computes an estimate of inputs as a function of future time on a logarithmically compressed scale and can be used to generate a scale-invariant power-law-discounted estimate of expected future reward. The representation of future time retains information about what will happen when. The entire timeline can be constructed in a single parallel operation that generates concrete behavioral and neural predictions. This computational mechanism could be incorporated into future reinforcement learning algorithms.

Neural Computation **31, 681–709** (2019)
doi:10.1162/neco_a_01171

1 Introduction

The ability to learn and operate in a continuously changing world with complex temporal relationships is critical for survival. For example, rats have to navigate around narrow holes and across wide fields; they have to learn that some stimuli present imminent danger requiring quick action, while others can serve as cues for events that will take place in a more distant future. Understanding the neural mechanisms that govern such behavioral flexibility and building artificial agents that have such capacity pose significant challenges for neuroscience and artificial intelligence.

In reinforcement learning (RL), an agent learns how to optimize its actions from interacting with the environment. The traditional approach to RL is to consider each different configuration of the environment as a different state (Sutton & Barto, 1998). Temporal difference (TD) learning has been employed to learn the scalar value of temporally discounted expected future reward for each state. This approach has been tremendously useful and has led to numerous practical applications (Mnih et al., 2015).

In this letter, we introduce a method for computing an estimate of future events along a logarithmically compressed timeline—an estimate of what will happen when in the future. This method addresses two major limitations of mainstream RL algorithms. First, because TD learning attempts to estimate an integral over a function of future time, it discards detailed information about the time at which future events are expected to take place. Of course, human decision makers can reason about the time at which future events will occur, leading many authors to augment the fast value computation supplied by TD learning with a model-based system (see Daw & Dayan, 2014, for a review). The model-based system is typically assumed to be slow; for some standard algorithms, the time taken to predict an outcome n steps in the future requires n matrix operations. Second, because the goal of TD learning is to estimate the exponentially discounted expected cumulative future reward, the method necessarily introduces a characteristic timescale (see Figure 1a). If the delay associated with the to-be-learned relationship is small compared to this scale, the behavior of the model will be dramatically different than if it is large compared to this scale.[1] In this letter, we present an alternative method for predicting future outcomes in continuous time that addresses these limitations.

1.1 Fixing a Timescale Limits Flexibility. Consider the task of designing an agent that will be deployed in a realistic environment without additional intervention from the designer. Successful performance on many tasks requires the ability to learn across a range of timescales. To make this example more concrete, consider designing an agent that will be deployed

[1] Similar arguments can be made when eligibility traces are considered.

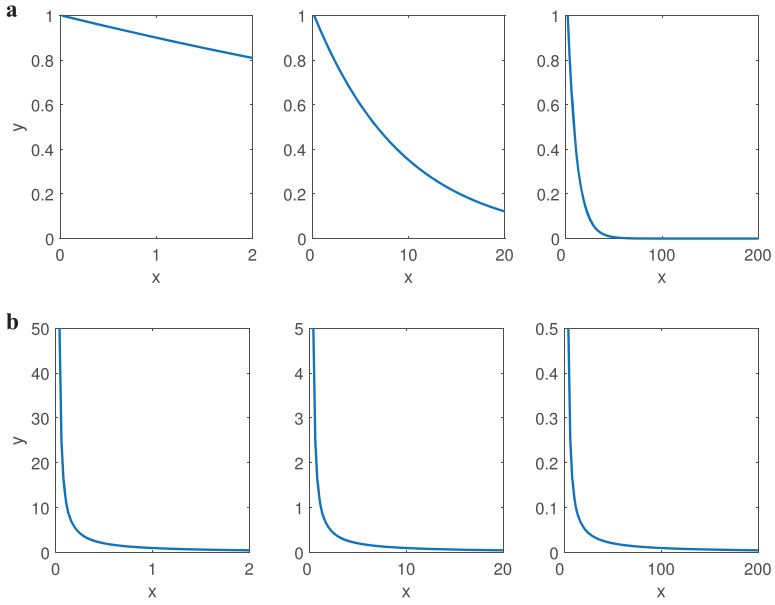

Figure 1: Exponential discounting introduces a scale; power law discounting is scale free. (a) An exponential function has qualitatively different properties at different scales. The function $y = \gamma^x$ is shown at three different scales for $\gamma = 0.9$. If x is on the order of the time constant $(-\log \gamma)$ we obtain the familiar exponential function with a clearly defined gradient for different values of x (middle). If x is small with respect to the time constant, we find a linear function with a shallow slope (left). If x is large relative to the time constant (right), the function approximates a delta function with a peak around zero. (b) Power law discounting ($y = x^{-1}$). For all ranges of x values, the power law gives the same relative gradient of values.

on the streets of Boston to learn to complete the everyday tasks of a postdoc. In order to get from Boston University to Harvard, the agent must learn that switching onto the Red Line leads to Harvard Square about 20 minutes in the future. At Dunkin Donuts, the agent must learn that paying money leads to a cup of coffee in about 1 minute. Grasping the cup and sipping the coffee predicts the taste of coffee immediately but also predicts the stimulating effect of caffeine several minutes in the future. In designing an agent to learn all of these tasks in an unknown environment, the designer will not necessarily know what temporal scales are important. We thus desire that the learning algorithm be scale invariant.

Algorithms based on the Bellman equation, which includes TD learning, estimate an exponentially discounted expected future return (value) by harnessing the recursive structure of the value function $V(t)$,

$$V(t) = \mathbb{E}[r(t) + \gamma V(t+1)], \tag{2.1}$$

where $r(t)$ denotes the reward at time t, the expectation represents an average over future events, and the exponential discount factor γ fixes a characteristic timescale.[2] The scaling results in very different policies at different temporal scales (see Figure 1a). Consider a world in which two rewards, A and B, follow a cue. The delay from the cue to B is twice the delay to A. Suppose that we do not know the units of time in the world and pick $\gamma = 0.9$. If the units of the world are such that the delay to A is 1 and the delay to B is 2, then the agent would prefer B if the value of A was \$1 and B was \$1.20. However, if the units of the world are very different such that the delay to A was 100 and the delay to B was 200, then even if the reward at B was \$30,000, the agent would still prefer A. This example makes clear that the success of a model that makes use of exponential discounting depends critically on aligning the choice of γ to the relevant scale of the world. In addition, animal literature suggests that hyperbolic discounting explains the data better than exponential discounting (see, e.g., Green & Myerson, 1996), for instance, regarding preference reversal (Green & Myerson, 2004; Hayden, 2016).

1.2 Representing the Future with a Scalar Obscures Temporal Information. One could implement scale-invariant power-law discounting[3] by choosing an appropriate spectrum of exponential discount rates (Kurth-Nelson & Redish, 2009; Sutton, 1995). However it is computed, a discounted value discards potentially important information about when an anticipated event will occur. For instance, consider the decision facing an agent about whether to buy a cup of very hot coffee. Drinking the coffee immediately would burn one's mouth. However, drinking the coffee after waiting a few minutes for it to cool down will result in a delicious and stimulating beverage. Is the value of the coffee negative (burned mouth) or positive (delicious beverage), or some weighted sum of the two? One way to answer the question is to state that the value of the coffee is a function over future time that is initially negative and then later positive. If the only information about this function that can be brought to bear in deciding whether to purchase the coffee is a single scalar value, then the decision maker may choose an inappropriate action, either purchasing the coffee when she does not have time to wait for it to cool or missing the opportunity to enjoy a delicious beverage in the near future.

One could tackle this problem using model-free RL approaches by expanding the state space to include relevant variables, such as the elapsed

[2] More precisely, the inverse of the time constant goes like $-\log \gamma$.

[3] If $f(t) = t^a$, then rescaling the time axis preserves the relative values at all time points, $f(\alpha t) = C f(t)$.

time that the cup has been held and the steam coming from the cup. However, when the elapsed time is one of the variables that the agent needs to keep track of, this approach becomes computationally very expensive. This is because the number of states rapidly increases. If time is discretized into m bins and we need to keep track of n stimuli, then the number of states is m^n. This is especially costly when time is discretized in equal-sized bins as in complete serial compound representation. Using microstates characterized with a set of compressed temporal basis functions as in Ludvig, Sutton, and Kehoe (2008, 2012) reduces the number of states to some degree, but this type of representation does not provide a future timeline.

Classical model-based RL enables decisions that take into account the time at which future events will take place. However, the computational cost of traditional model-based solutions grows linearly with the horizon over which one needs to estimate the future. In this letter, we present a method that constructs a function over future time for each stimulus (state). This representation of the future is logarithmically compressed, and the estimate of the future at many different points in time can be computed in parallel. One could compute an integral over this representation to maintain a cached value with power law discounting. But because the entire function is available, an agent can also incorporate the time at which rewards will become available into its decision making.

1.3 Scale-Invariant Temporal Representations in the Brain. The basic computational strategy we pursue is to (1) compute a scale-invariant representation of the temporal history leading up to the present and (2) at each moment, associate the history with the stimulus observed in the present. Step 1 assumes the existence of a scale-invariant compressed representation of temporal history. Step 2 assumes the existence of an associative mechanism. There is ample neural evidence for both of these assumptions. A large body of literature from cellular neuroscience provides evidence for an associative mechanism implementing Hebbian plasticity at synapses (Bliss & Collingridge, 1993; Lisman, Schulman, & Cline, 2002), which would be required for step 2.

There is also a growing body of evidence consistent with assumptions necessary for step 1. Experiments from several species suggest that the brain maintains a compressed representation of time in multiple brain regions. "Time cells" fire during a circumscribed period of time within a delay interval (Pastalkova, Itskov, Amarasingham, & Buzsaki, 2008; MacDonald, Lepage, Eden, & Eichenbaum, 2011); a reliable sequence of time cells tiles the delay on each trial (see Figure 5a). Because the sequence is reliable, time cells can be used to reconstruct how long in the past the delay began. In many experiments, these sequences also carry information about what stimulus initiated the delay interval (Pastalkova et al., 2008; MacDonald, Carrow, Place, & Eichenbaum, 2013; Tiganj, Cromer, Roy, Miller, & Howard, 2018; Terada, Sakurai, Nakahara, & Fujisawa, 2017). Because there are fewer cells that

fire later in the sequence and those that fire later in the sequence fire for a longer duration (Howard, Shankar, Aue, & Criss, 2015; Salz et al., 2016), the ability to reconstruct time decreases as the start of the interval recedes into the past. Time cells have been observed in several brain regions, including hippocampus (MacDonald et al., 2011; Salz et al., 2016), prefrontal cortex (Tiganj, Kim, Jung, & Howard, 2016; Bolkan et al., 2017; Tiganj et al., 2018), and striatum (Mello, Soares, & Paton, 2015; Akhlaghpour et al., 2016), in several species (Mau et al., 2018; Adler et al., 2012; Tiganj et al., 2018) and in a wide variety of behavioral tasks.

Taken together, these data indicate that at each moment, the brain maintains a temporal record of what happened when leading up to the present. The decrease in accuracy for events further in the past suggests that this temporal record is compressed. As such, these neural data align with long-standing predictions from cognitive models (Brown, Neath, & Chater, 2007; Balsam & Gallistel, 2009; Howard et al., 2015). These models further predict that the form of compression should be logarithmic. Behavioral models built from a logarithmically compressed representation readily account for scale-invariant behavior (Howard et al., 2015).[4]

1.4 Overview of This Letter. In this letter, we use a logarithmically compressed record of the past—a set of appropriate time cells—to construct a scale-invariant estimate of the time of future events. A logarithmically compressed record of the past can be efficiently computed using a method we describe in detail below (Shankar & Howard, 2012, 2013). At each moment, this representation of the past is associated to the present. Neurally, this association requires nothing more elaborate than Hebbian plasticity, which can be implemented via long-term potentiation (Bliss & Collingridge, 1993). The past-to-present association can also be understood as a present-to-future association. As such, multiplying this association with the present stimulus vector enables us to identify the sequence of stimuli that will follow the probe stimulus at different points in the future. Section 2 describes this method more precisely.

This method yields an estimate of the future that has very different properties from traditional approaches used in RL. The properties of this representation are described with illustrative examples in section 3. Because the representation of the past is logarithmically compressed, so too is the estimate of the future that it produces. A cached scalar value can be computed from this timeline, yielding (scale-invariant) power law discounting by summing over the predicted future. Notably, the compressed timeline representation also provides a function over simulated time. The future

[4] For much the same reason that, on a logarithmic scale, the difference between 1 and 2 is the same as the difference between 100 and 200, models built from a logarithmically compressed temporal representation will be scale invariant.

timeline can be computed in a single parallel operation and sums over potential outcomes. Section 4 describes neural and behavioral predictions of the model, reviewing recent empirical results that are consistent with the proposed hypothesis that the brain constructs a logarithmically compressed future.

2 Constructing a Logarithmically Compressed Timeline of the Future

This approach requires two key components: a logarithmically compressed memory representation and an associative memory between the compressed representation and the present stimulus. Section 2.1 describes a method for constructing a logarithmically compressed memory representation following Shankar and Howard (2013). Section 2.2 describes the associative memory. Section 2.3 describes the future timeline that results from probing the associative memory with a stimulus representation.

2.1 Previous Work: Constructing a Compressed Memory Representation of the Past. Consider a case in which the network is presented with a vector-valued input that changes over time $\mathbf{f}(t)$. This input reflects the presence or absence of a set of discrete stimuli (states) that we denote as $I = \alpha,\ \beta,\ \gamma \ldots$ For simplicity, let us assume that the input uses a localist (one-hot) representation; if stimulus α is present at time t, we write $f_\alpha(t) = 1$. The goal of this method is to construct an estimate of the past leading up to the present. We refer to this memory representation as $\tilde{\mathbf{f}}$. A temporal record of the past requires two types of information. In order to estimate $\mathbf{f}(t' < t)$, we need to maintain both what and when information. Thus, we index each of the neurons in $\tilde{\mathbf{f}}$ by two indices, $\tilde{f}_{\overset{*}{\tau},i}$ (see Figure 2). The second index $i \in I$ corresponds to the what information. The other index, $\overset{*}{\tau}$, refers to the time in the past that this neuron is attempting to represent. That is, the network includes a set of values of $\overset{*}{\tau} \in \{\overset{*}{\tau}_1,\ \overset{*}{\tau}_2,\ \overset{*}{\tau}_3 \ldots\}$. Because the value of $\overset{*}{\tau}$ for the ith row of the network, $\overset{*}{\tau}_i$, has physical meaning, we refer to the neurons in $\tilde{\mathbf{f}}$ by their value of $\overset{*}{\tau}$ rather than their row number. Each entry $\tilde{f}_{\overset{*}{\tau},\alpha}(t)$ approximates $f_\alpha(t + \overset{*}{\tau})$. Here the values of $\overset{*}{\tau}$ are negative as they refer to a temporal distance in the past relative to the present.

Following prior work (Shankar & Howard, 2012, 2013), we will construct the representation of the past $\tilde{\mathbf{f}}$ by means of an intermediate representation $\mathbf{F}$. Each neuron in $\mathbf{F}$ aligns with a corresponding neuron in $\tilde{\mathbf{f}}$ (see Figure 3a). The neurons in $\mathbf{F}$ are indexed by the label of the stimulus in the world that activates them ($\alpha,\ \beta,\ \gamma \ldots$) and a scalar value s. The values of s for each row of $\mathbf{F}$ align with the corresponding values of $\overset{*}{\tau}$ in each row of $\tilde{\mathbf{f}}$ (see Figure 3b):

$$F_{s,\alpha} \leftrightarrow \tilde{f}_{\overset{*}{\tau},\alpha}. \tag{2.1}$$

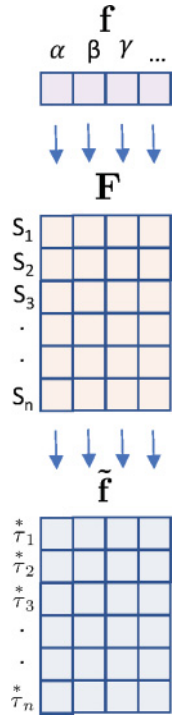

Figure 2: Constructing a memory representation of the recent past. The schematic shows the two-layer network for constructing a memory representation by implementing an approximation of the Laplace and the inverse Laplace transform. The input **f** is a vector over states α, β, γ. . . This provides input to a two-layer network where each layer is a 2D array (sheet) of neurons. Neurons in the first layer **F** are leaky integrators indexed by the state they encode, α, β, γ. . . and their rate constant s. We refer to the activation of a particular entry as $F_{s,\alpha}$. Neurons in the second layer $\tilde{\mathbf{f}}$ activate sequentially following the input stimulus. They are indexed by the state that provides their input and the time by which the peak of their activation follows the input stimulus $\overset{*}{\tau}$. The activation of a particular unit is referred to as $\tilde{f}_{\overset{*}{\tau},\alpha}$.

The mapping between s and $\overset{*}{\tau}$ is such that $s = -k/\overset{*}{\tau}$, where k is an integer with physical meaning that we will describe and $i \in 1, 2, 3 \ldots n$, where n is the number of rows in **F** and $\tilde{\mathbf{f}}$. As with $\overset{*}{\tau}$, there are a finite set of values of s, $s \in \{s_1,\ s_2,\ s_3 \ldots\}$. As with $\overset{*}{\tau}$, there is a physical meaning to the ith value of s so we refer to neurons in **F** by their value of s_i rather than their index i. Values of s are defined to be positive. Following previous work (Shankar &

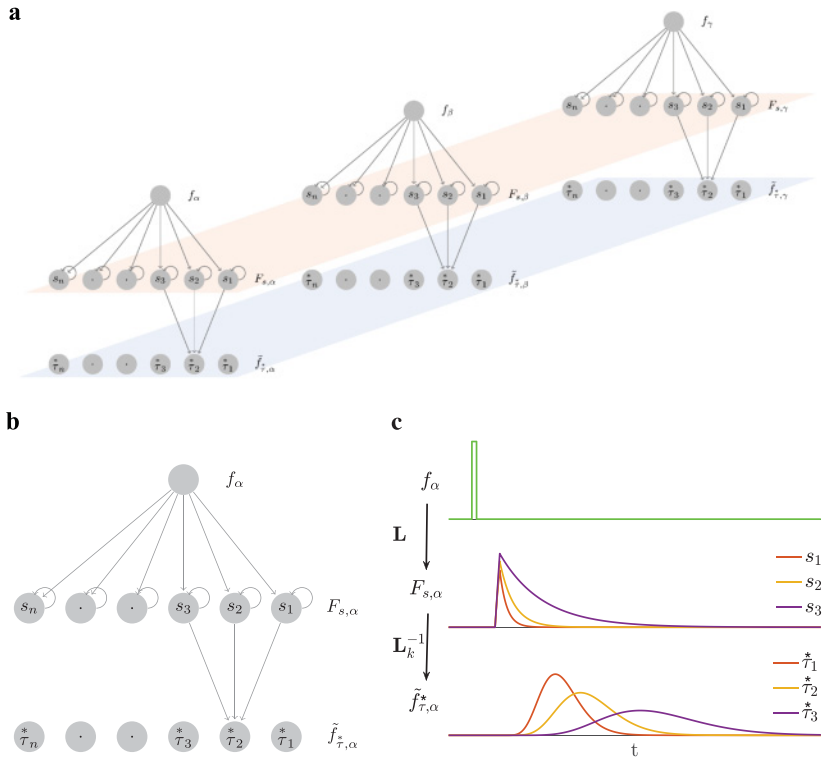

Figure 3: Structure and dynamics of the memory representation. (a) The two-layer network is organized such that each input state has its own set of units in the $\mathbf{F}$ and $\tilde{\mathbf{f}}$ layers (when constructing the memory representation, there is no crosstalk between the neurons that correspond to different states). (b) The input stimulus $f_\alpha(t)$ feeds into a layer of leaky integrators $F_{s,\alpha}$ that implement a discrete approximation of an integral transform. Each neuron in the first layer has a characteristic rate constant s_i. $F_{s,\alpha}$ projects onto $\tilde{f}^{*}_{\tau,\alpha}$ through a set of weights defined with the operator $\mathbf{L}_k^{-1}$, which implements an approximation of the inverse of the Laplace transform. Notice that the $\mathbf{L}_k^{-1}$ operator projects only to a local neighborhood (k neurons). Each of the neurons in the second layer has its characteristic peak time relative to the input onset $\overset{*}{\tau}_i$. The analytic relationship between $\overset{*}{\tau}$ and s can be expressed as $\overset{*}{\tau} = -k/s$. Thus, choosing $\overset{*}{\tau}$ and integer k fully specifies s; similarly, choosing s and k fully specifies $\overset{*}{\tau}$. We chose $\overset{*}{\tau}$ to be logarithmically spaced (in order to have a logarithmically compressed memory representation). (c) A response of the network to a delta function input. Activity of only three neurons in each layer is shown. Neurons in $\tilde{f}^{*}_{\tau,\alpha}$ activate sequentially following the stimulus presentation. The width of the activation of each neuron scales with the peak time determined by the corresponding $\overset{*}{\tau}$, making the memory scale-invariant.

Howard, 2013; Howard & Shankar, 2018), we choose the values of $\overset{*}{\tau}$ and s to be evenly spaced on a logarithmic scale.[5]

The dynamics of each unit in $\mathbf{F}$ obeys

$$\frac{dF_{s,\alpha}(t)}{dt} = -sF_{s,\alpha}(t) + f_\alpha(t), \tag{2.2}$$

where the value of s on the right-hand side refers to that particular neuron's value s_i. Here we can see that s describes each neuron's rate constant; $1/s$ describes each neuron's time constant. Taking the network state across all values of s, $\mathbf{F}(s)$ estimates the Laplace transform of $\mathbf{f}(t' < t)$. To see that $F_{s,\alpha}$ at time t is the Laplace transform of $f_\alpha(t' < t)$, solve equation 2.2:

$$F_{s,\alpha}(t) = \int_{-\infty}^{t} e^{-s(t-t')} f_\alpha(t')dt'. \tag{2.3}$$

Knowing that $\mathbf{F}$ at time t holds the Laplace transform of $\mathbf{f}$ leading up to the present suggests a strategy to construct an estimate of $\mathbf{f}$. If we could invert the transform and write the answer into another set of neurons $\tilde{\mathbf{f}}$, this would provide an estimate of $\mathbf{f}$ as a function of time leading up to the present. The Post approximation (Post, 1930) provides a recipe for approximating the inverse transform that can be computed with a set of feedforward weights, which we denote $\mathbf{L}_k^{-1}$:

$$\tilde{f}_{\overset{*}{\tau},\alpha}(t) = \mathbf{L}_k^{-1} F_{s,\alpha}(t). \tag{2.4}$$

The integer k determines the precision of the approximation. Denoting the kth derivative with respect to s as $F_{s,\alpha}^{(k)}$, we can rewrite equation 2.4 as

$$\tilde{f}_{\overset{*}{\tau},\alpha}(t) = C_k s^{k+1} F_{s,\alpha}^{(k)}(t), \tag{2.5}$$

where C_k is a constant that depends only on k.

To get an intuition into the properties of $\tilde{f}$, we present a delta function to f_α at time zero and examine the activity of $F_{s,\alpha}(t)$ and $\tilde{f}_{\overset{*}{\tau},\alpha}(t)$. We find immediately that $F_{s,\alpha}(t) = e^{-st}$. Moreover, the activity of the neurons in $\tilde{\mathbf{f}}$ obeys

$$\tilde{f}_{\overset{*}{\tau},\alpha}(t) = C_k \frac{1}{\overset{*}{\tau}} \left(\frac{t}{\overset{*}{\tau}}\right)^k e^{-k\frac{t}{\overset{*}{\tau}}}, \tag{2.6}$$

[5] For instance, one can choose $\overset{*}{\tau}_i = \overset{*}{\tau}_{\min}(1+c)^{i-1}$ for some minimum value of $\overset{*}{\tau}$, $\overset{*}{\tau}_{\min}$ and a constant c that controls the spacing.

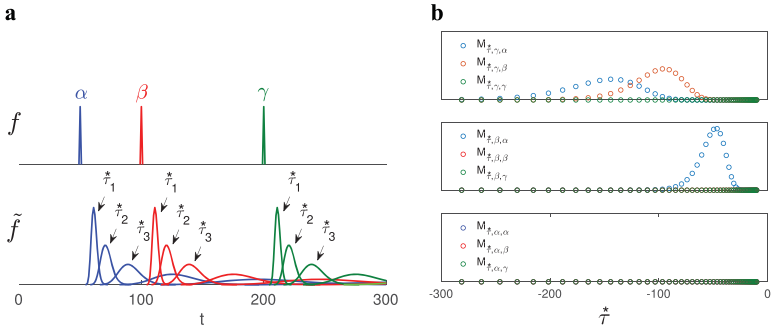

Figure 4: Constructing an associative memory by building connections between present inputs and memory of the recent past. (a) Illustration of the compressed memory representation $\tilde{\mathbf{f}}$ as a function of time during presentation of the sequence α, β, γ. Stimuli presented at different times (top) induce sequential activation (bottom) in $\tilde{\mathbf{f}}$. Activation corresponding to different stimuli is shown with different colors. For clarity, only a handful of neurons are displayed. (b) A graphical depiction of the state of the associative memory $\mathbf{M}$ after learning of the sequence α, β, γ. Each of the three plots shows entries stored in $\mathbf{M}$ as a function of τ^*. $\mathbf{M}_{\tau^*,q,r}$ is the synaptic weight from r to q at a particular τ^*. The different stimuli are shown in different colors in each plot. Top: Associations stored in $M_{\tau^*,\gamma,\alpha}$, $M_{\tau^*,\gamma,\beta}$, and $M_{\tau^*,\gamma,\gamma}$ as functions of log-spaced τ^*. Because γ was preceded by both α and β, both $M_{\tau^*,\gamma,\alpha}$ and $M_{\tau^*,\gamma,\beta}$ have peaks. Because β had been presented more recently when γ was presented, the curve for $M_{\tau^*,\gamma,\beta}$ has a peak closer to $\tau^* = 0$. Because the representation of times in the more recent past is more accurate than times further in the past, the peak for β is also sharper as a function of τ^*. Middle: Associations stored in $M_{\tau^*,\beta,\alpha}$, $M_{\tau^*,\beta,\beta}$, and $M_{\tau^*,\beta,\gamma}$. Because β was preceded by α at a short lag, $M_{\tau^*,\beta,\alpha}$ differs from zero at low values of τ^*. Since β was not preceded by itself or by γ, blue and green traces are zero. Bottom: Associations stored in $M_{\tau^*,\alpha,\alpha}$, $M_{\tau^*,\alpha,\beta}$, and $M_{\tau^*,\alpha,\gamma}$. Since α was not preceded by any stimulus, corresponding entries in $\mathbf{M}$ are zero for all values of τ^*.

where C_k here is a different constant that depends only on k. The activity of each node in $\tilde{f}_{\tau^*,\alpha}$ is the product of an increasing power term $\left(\frac{t}{\tau^*}\right)^k$ and a decreasing exponential term $e^{-k\frac{t}{\tau^*}}$. In the time following a delta function input, the firing of each neuron in $\tilde{f}_{\tau^*,\alpha}$ peaks at τ^* (see Figure 3c). Thus, following a transient input of state α, neurons in $\tilde{f}_{\tau^*,\alpha}$ activate sequentially.

Figure 4a shows the sequential, spreading activation with logarithmically spaced τ^* for three different transient stimuli. This mathematical model for estimating the past has properties that resemble sequentially activated time cells (compare to Figure 5; see also Howard et al., 2014; Tiganj et al.,

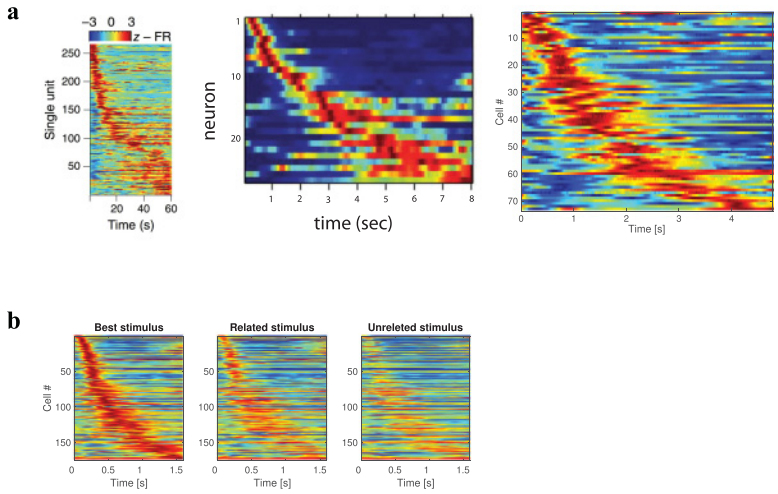

Figure 5: Recordings of sequentially activated time cells in different behavioral tasks, different mammalian species, and different brain regions. (a) Sequences of time cells in the brain contain information about the time of past events. Each of the three plots shows the activity of many neurons during the delay period of a behavioral task. In each plot, each row gives the average firing rate as a function of time for one neuron. Red colors indicate a high firing rate. Because neurons fire for a circumscribed period of time during the delay, these neurons could be used to decode the time at which the delay started. Put another way, each neuron can be understood as coding for the presence of the start of the delay at a lag $\overset{*}{\tau}$ in the past. The number of cells active at any one time decreases as the delay unfolds (note the curvature) and the firing fields spread (note the increasing width of the central ridge). This reflects a decrease in accuracy for time as the start of the delay recedes further into the past. From left to right: mouse mPFC during a spatial working memory task, after Bolkan et al. (2017); rat hippocampus, during the delay period of a working memory task, after MacDonald et al. (2011); rat mPFC during a delay period of temporal discrimination task, after Tiganj et al. (2016). (b) Sequentially activated cells in monkey lPFC encode time conjunctively with stimulus identity during delayed-match-to-category task. Animals were presented with stimuli chosen from four categories (dogs, cats, sports cars and sedan cars). Based on visual similarity the stimuli belonged to two category sets (animals and cars). The time interval shown on the plots includes a 0.6 s sample stimulus presentation and a 1 s delay interval that followed the sample stimulus. Each of the three heatmaps shows the response of every unit classified as a time cell. The units show distinct firing rate for different stimuli that started the delay interval, reflecting the visual similarity (the magnitude of the response for related stimulus was larger than for unrelated stimulus) and indicating stimulus selectivity of time cells. After Tiganj et al. (2018).

2018). Previous biophysical modeling has developed a neurally plausible mechanism for implementing leaky integrators with a spectrum of time constants (Tiganj, Hasselmo, & Howard, 2015) and for constructing a circuit implementing the inverse transform (Liu, Tiganj, Hasselmo, & Howard, 2018).

$\tilde{\mathbf{f}}$ approximates $\mathbf{f}$ leading up to the present. However, the precision of the approximation decreases for events further in the past. One way to see this is that the duration over which $\tilde{f}_{\overset{*}{\tau},\alpha}$ is activated by a delta function input increases as one chooses larger values of $|\overset{*}{\tau}|$. However, this inaccuracy is scale-invariant; the spread in time for a neuron with a particular $\overset{*}{\tau}$ is a rescaled version of the firing of another neuron that received the same input but has a different value of $\overset{*}{\tau}$. Put another way, the activity of every neuron receiving a delta function input obeys the same time dependence in units of $t/|\overset{*}{\tau}|$. This rescaling of the activity of neural response in time also has a correspondence in the pattern of activity across neurons with different values of $\overset{*}{\tau}$ as the stimulus recedes into the past. At any moment, when the stimulus is t_o time in the past, there is a bump of activity centered around the neurons with $\overset{*}{\tau} \simeq t_o$. However, the difference in the value of $\overset{*}{\tau}$ between adjacent neurons is not constant (e.g., note the increasingly spread points in Figure 4b). With logarithmic spacing of $\overset{*}{\tau}$ values, the shape of the bump of activity across neuron number remains of constant width as the stimulus recedes into the past (Howard et al., 2015).

2.2 Constructing an Associative Memory. At each time t, an associative memory tensor $\mathbf{M}$ is updated with the outer product of the current input state $\mathbf{f}$ and $\tilde{\mathbf{f}}$ (see Figure 4b). Hence, $\mathbf{M}$ is a three-tensor. At each moment, $\mathbf{M}$ is updated with the simple Hebbian learning rule:

$$\frac{dM_{\overset{*}{\tau},\beta,\alpha}}{dt} = \lambda\, f_\beta(t)\, \tilde{f}_{\overset{*}{\tau},\alpha}(t). \tag{2.7}$$

Here λ is a learning rate that we choose to be 1. $\mathbf{M}$ can be implemented as a set of synaptic weights learned through Hebbian plasticity. Because $\tilde{\mathbf{f}}_{\overset{*}{\tau},\alpha}$ stores a coarsely grained estimate of the past, its average over many experiences, $\mathbf{M}$, is a coarsely grained estimate of the lagged cooccurrence of each pair of states,

$$M_{\overset{*}{\tau},\beta,\alpha} \propto P\left[\mathbf{f}(t + |\overset{*}{\tau}|) = \beta,\ \ \mathbf{f}(t) = \alpha\right], \tag{2.8}$$

where P denotes probability.

We can also construct an estimate of the conditional probability by normalizing $\mathbf{M}$ as follows:

$$\bar{M}_{\overset{*}{\tau},\beta,\alpha} \equiv \frac{M_{\overset{*}{\tau},\beta,\alpha}}{\sum_{i\in I} M_{\overset{*}{\tau},\beta,i}}. \tag{2.9}$$

One could imagine that this normalization is implemented online by a divisive presynaptic normalization mechanism (Beck, Latham, & Pouget, 2011). Now $\bar{\mathbf{M}}$ is an associative memory that provides a coarsely grained estimate of the conditional probability of state β following state α at a lag of $|\overset{*}{\tau}|$:

$$\bar{M}_{\overset{*}{\tau},\beta,\alpha} \propto P\left[\mathbf{f}(t + |\overset{*}{\tau}|) = \beta \mid \mathbf{f}(t) = \alpha\right]. \tag{2.10}$$

As we will see in the section 2.3, by multiplying $\bar{\mathbf{M}}$ from the right with a current state, we can generate the probability of all other states following at each possible lag.

2.3 Estimating a Future Timeline. $\mathbf{M}$ stores the pairwise temporal relationships between all stimuli subject to logarithmic compression. At the moment a state is experienced, the history leading up to that state is stored in $\mathbf{M}$ (see equation 2.7). After many presentations, $\mathbf{M}$ records the probability that each state is preceded by every other state at each possible lag. This record of the past can also be used to predict the future. By multiplying $\bar{\mathbf{M}}$ with the current state from the right, we can generate an estimate of the future. In a general case, let us consider $\mathbf{f}(t)$, which can have multiple stimuli presented at the same time. Stimuli that will follow the present input $\mathbf{f}(t)$ at a time lag $|\overset{*}{\tau}|$ can be estimated from the information recorded in $\bar{\mathbf{M}}$:

$$\mathbf{p}_{-\overset{*}{\tau}} \equiv \bar{\mathbf{M}}_{\overset{*}{\tau}}\mathbf{f}, \tag{2.11}$$

$$p_{-\overset{*}{\tau},\beta} = \sum_{i\in I} \bar{M}_{\overset{*}{\tau},\beta,i} f_i. \tag{2.12}$$

Like $\mathbf{F}$ and $\tilde{\mathbf{f}}$, $\mathbf{p}$ can be understood as a 2D array indexed by stimulus identity and $\overset{*}{\tau}$. However, whereas for $\tilde{\mathbf{f}}$, $\overset{*}{\tau}$ is negative corresponding to estimates of the past, for $\mathbf{p}$ the values of $\overset{*}{\tau}$ are positive, corresponding to estimates of the future. The value of $\overset{*}{\tau}$ for the ith row of $\tilde{\mathbf{f}}$ and the value of $\overset{*}{\tau}$ for the ith row of $\mathbf{p}$ have the same magnitude but are opposite in sign. $p_{-\overset{*}{\tau},\beta}$ is a magnitude of the prediction that state β will follow present input $\mathbf{f}$ at a time lag $|\overset{*}{\tau}|$. When the input is interpretable as a probability density function (when $\sum|\mathbf{f}| = 1$), then $\mathbf{p}_{-\overset{*}{\tau}}$ is also a probability density function. When $\mathbf{f}$ is not a probability density function, $\mathbf{p}_{-\overset{*}{\tau}}$ is not either.

In a more specific case, when $\mathbf{f}(t)$ can have only one stimulus presented at the same time, the magnitude of the prediction that state β will follow the present input, say, state α, at a time lag $|\overset{*}{\tau}|$ is a scalar stored in $\bar{M}_{\overset{*}{\tau},\beta,\alpha}$:

$$p^{\alpha}_{-\overset{*}{\tau},\beta} = \bar{M}_{\overset{*}{\tau},\beta,\alpha}. \tag{2.13}$$

Note that $\mathbf{p}^{\alpha}$ inherits the same compression present in $\tilde{\mathbf{f}}$. The "blur" in the estimate of the time of presentation of a past stimulus in $\tilde{\mathbf{f}}$ with $\overset{*}{\tau} < 0$ naturally leads to an analogous blur in $\mathbf{p}^{\alpha}$ as a function of future time $\overset{*}{\tau} > 0$. Expected future outcome at a lag $\overset{*}{\tau}$ can be estimated by examining the states predicted at that lag and estimating the reward status of each. Properties of this representation of future time are illustrated in more detail in section 3.

3 Illustrating the Properties of the Representation of Future Time

In this section we illustrate properties of the representation of future time constructed by multiplying $\bar{\mathbf{M}}$ with a particular state vector (see equation 2.13). In section 3.1, we demonstrate that the representation that results is scale-invariant. In section 3.2, we show that a cached value for each state can be computed, resulting in a scale-invariant value that is discounted according to a power law. In section 3.3, we illustrate the flexibility of this method in generating nonmonotonic functions enabling the user to solve problems such as the hot coffee problem described in section 1. In section 3.4, we demonstrate that future time gives an estimate summed over all possible paths. Finally, in section 3.5, we demonstrate application of this approach in decision making.

3.1 Scale Invariance of Future Time. If two environments differ only in their temporal scale, an artificial agent based on a scale-invariant algorithm will take the same actions in both environments. This property is illustrated for this method through a simple toy example in Figure 6. In this example, there are two states to choose from, α and β, and a third rewarding state R that the agent is interested in predicting. The two environments shown in Figure 6 differ only in the temporal spacing between different stimuli. The bottom environment (marked as Scale 4 in Figure 6b) is a temporally stretched version of the top environment (marked as Scale 1 in Figure 6a). Stretching the time axis of the top environment by four times would give exactly the bottom environment. At the decision point D at time $t = 0$, the agent needs to choose either state α or state β (the example is designed as a deterministic Markov decision process, so taking an action can be understood as directly selecting a state).

Under the assumption that the agent has explored the environment by choosing each direction at least once, all needed temporal associations are stored in $\bar{\mathbf{M}}$. The next time that the agent faces the decision point at time $t = 0$, it can construct the future time as in equation 2.13. The predictions $\mathbf{p}^{\alpha}$ and $\mathbf{p}^{\beta}$ constructed separately for α and β both give power law discounted estimates of the expected future outcome that rescale with rescaling of the environment.

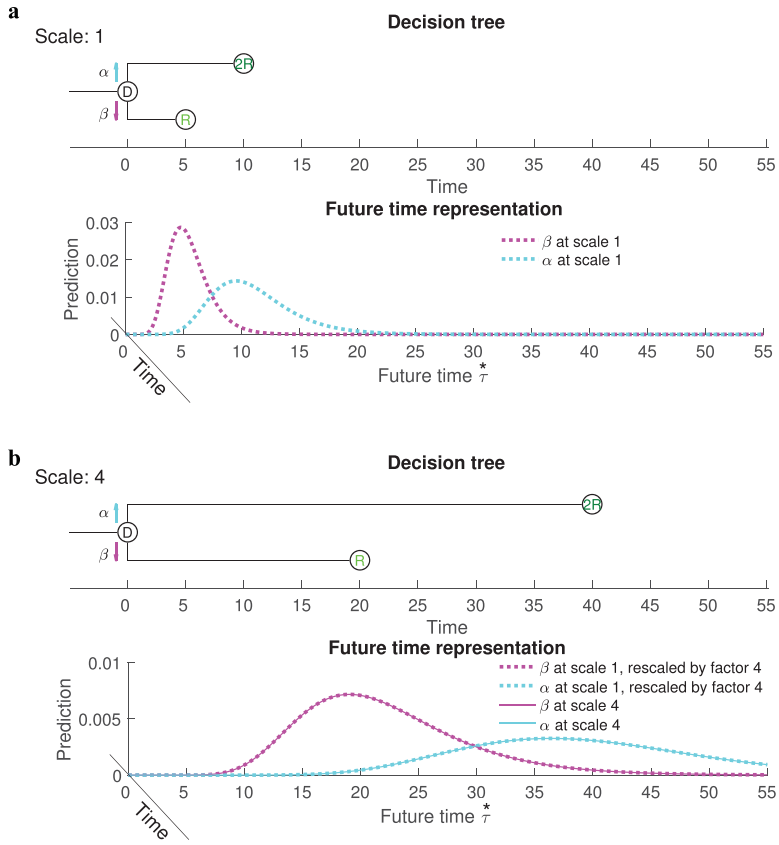

Figure 6: The representation of future time is scale invariant. In this and subsequent figures, the agent evaluates the degree to which each of two states, α and β, predicts a desired outcome R. The decision tree in the environment is shown in the top panel. The estimate of the future cued by each of the two states α and β as a function of future time τ^* is shown by the lines at the bottom. (a) A simple decision tree in which α predicts reward after 10 units of time and β predicts reward after 5 units of time. The dashed lines show that the cues predict reward at different times. Note that the prediction of events further in the future is made with less precision. (The reward predicted by β is twice the size of the reward predicted by α to make the figure easier to read.) (b) The same decision tree, but with the temporal intervals rescaled by a factor of 4. The solid lines show the predictions from the environment in panel a rescaled by a factor of 4 (i.e., stretched by a factor of 4 and multiplied by 4). Note that the functions in the two environments are precisely rescaled versions of one another.

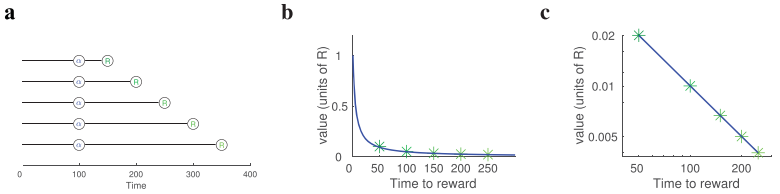

Figure 7: The value aggregated by integrating over future time obeys power law discounting. (a) Constructing prediction of the future reward. An agent observed a temporal sequence consisting of a state α, followed by a rewarding state R at some delay. (b) The value of α computed according to equation 3.1 as a function of the delay between α and reward (expressed in the units of the reward R). The values associated with the five evenly spaced delays in panel a are shown as star symbols. The blue line is a power law with exponent -1. (c) Same as panel b but on log-log axes.

3.2 Computing Cached Power-Law Discounted Stimulus Value by Integrating over the Timeline. There are circumstances where a decision maker does not have time to evaluate a compressed function over future time and a cached value of each state would be sufficient. A cached value can be computed by maintaining, for each state, an average value over future time updated by taking an integral over the future,

$$V^{\alpha}(t) = \sum_{i \in I} r_i \int_0^{\infty} p^{\alpha}_{\overset{*}{\tau},i}\, g_{\overset{*}{\tau}}\, d\overset{*}{\tau}, \tag{3.1}$$

where $\mathbf{r}$ is a column vector describing the value of each state and $g_{\overset{*}{\tau}}$ is the number density of $\overset{*}{\tau}$ values $\frac{dN}{d\overset{*}{\tau}}$. The number density specifies how many units are used to represent a particular spacing of $\overset{*}{\tau}$. For instance, if spacing between $\overset{*}{\tau}$ nodes would be linear, the number density $g_{\overset{*}{\tau}}$ would be 1. With logarithmic spacing of $\overset{*}{\tau}$, the number density goes down as $1/\overset{*}{\tau}$.

In order to ensure Weber-Fechner spacing, we set $g_{\overset{*}{\tau}} = 1/\overset{*}{\tau}$, but one could in general augment this by including a function to differentially weight the contribution of different values of $\overset{*}{\tau}$. As long as that function does not introduce a scale, the cached value computed in this way will remain scale invariant (power law). Figure 7 illustrates properties of value computed from equation 3.1.

Applying equation 3.1 to the example shown in Figure 6 reveals that the ratio of the values for states α and β is constant when time is rescaled. This means that the relative values assigned to various choices do not depend on the timescale of the environment, only on their relative magnitude and timing.

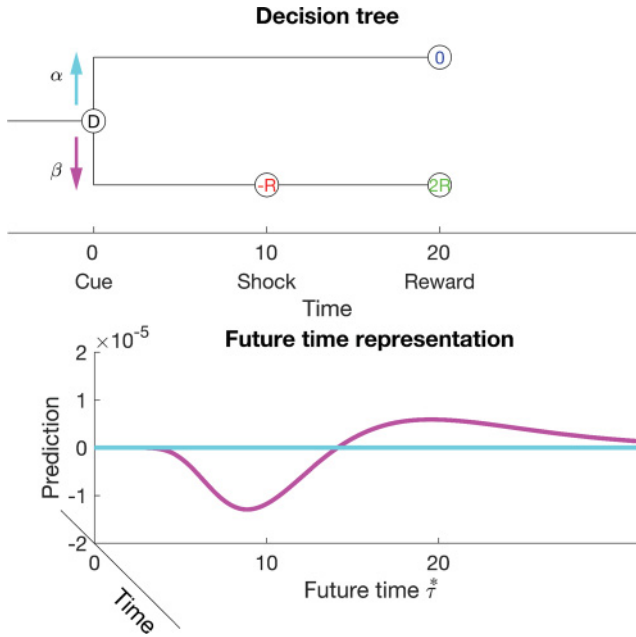

Figure 8: The estimate of a future timeline enables the decision maker to anticipate different outcomes at different points in time. (Top) Choice α is neutral, predicting neither reward nor punishment. Choice β results in a negative outcome (e.g., a shock) after 10 units of time and then a large positive outcome after 20 units of time. (Bottom) The representation of future time induced by each choice varies as a function of the temporal horizon. α is preferable to β at short delays, but β is preferable to α at longer delays. A decision maker could incorporate this information about the future time course when the choices are presented.

3.3 Nonmonotonic Functions over Future Time. In traditional RL, the value of each state is a scalar. The approach introduced here provides a recipe for simulating a function of a logarithmically compressed future. The example in Figure 8 illustrates one case in which this type of representation has an advantage over the scalar representation. In this example, state α is neutral; no meaningful outcome follows it. However, state β is followed sequentially by a negative outcome (e.g., a burned mouth) and then later by a positive outcome (e.g., delicious coffee). The ability to simulate outcomes as a function of future time can enable the agent to make decisions in a more flexible way (by dynamically altering the time horizon of planning) than would be possible if all the available information about the future was expressed as a scalar.

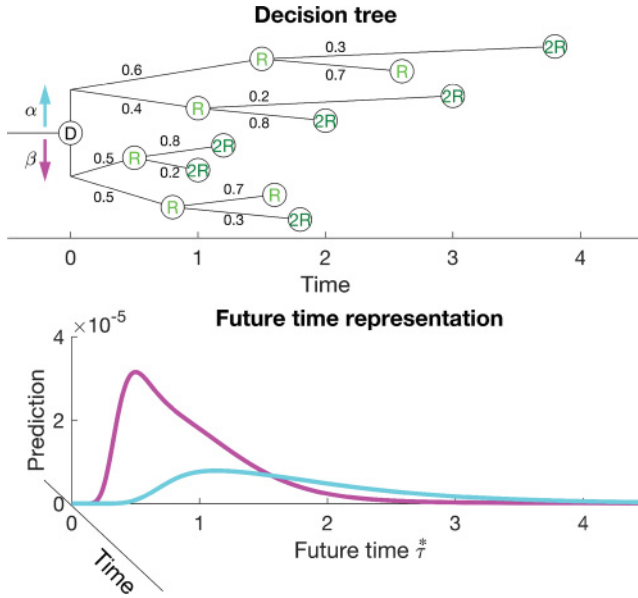

Figure 9: The estimate of a future timeline sums over all possible future trajectories. (Top) The decision tree for a complex choice. Both α and β lead to a complex set of possible outcomes that occur at specific times with given probabilities (shown as numbers near each branch of the tree). (Bottom) The estimate of a future timeline averages over different possible paths with each outcome weighted by its probability of occurrence given the choice stimulus (α or β).

Notice that the same amount of information is conveyed even when having only the set of exponentially decaying neurons (**F** neurons). However, applying the inverse Laplace transform and estimating the future as proposed here allows the agent to examine the future directly in the units of time, without need for an additional decoder. This type of representation provides direct access to temporal order and distance.

3.4 Future Time Sums over Trajectories. Figure 9 illustrates an important property of the proposed approach: simulated future time provides a probability of each stimulus a time $\overset{*}{\tau}$ in the future summed across all possible future trajectories. Let us assume that the agent has sampled the environment sufficiently many times to learn the transition probabilities and the temporal dynamics of the environment, which are now stored in $\tilde{\mathbf{M}}$. Now, computing the prediction $\mathbf{p}^{\alpha}_{\overset{*}{\tau}}$ provides an overall estimate of the reward averaged across all the future trajectories. However, it retains information about how far in the future those outcomes will be obtained. This property

allows a rapid evaluation of different decision trees. Evaluating a particular sequence of outcomes that depend on sequential actions would still require supplementing this representation with a more traditional model-based approach. Moreover, correctly learning the outcomes requires sampling the entire tree, which may be much slower than TD-based learning in an environment with Markov statistics.

Notice that in many problems in RL, states that follow the present state often change in response to the action taken by the agent. For simplicity we are studying the Pavlovian case (similar to previous authors like Schultz, Dayan, & Montague, 1997). In the control setting, we would need to simultaneously estimate $\tilde{\mathbf{M}}$ and a policy, since these are coupled. We think the interplay between prediction and control is important and leave that to future work.

3.5 Temporally Flexible Decision Making. The ability to construct a timeline of the future events enables flexible decision making that incorporates the decision maker's temporal constraints. For instance, consider making a decision about what to get for lunch while waiting for a train. The food option one pursues may be very different if one has 15 minutes before the train arrives than if one has an hour before the train arrives. Because the model carries separate information about when outcomes will be available, as well as their identity, it is possible to make decisions that differentially weight outcomes at different points in the future. If the decision maker has a temporal window over which outcomes are valuable, $w_{\tilde{\tau}^*}$, then one can readily compute value using a generalization of equation 3.1:

$$V^{\alpha}(t) = \sum_{i \in I} r_i \int_0^{\infty} p^{\alpha}_{\tilde{\tau}^*,i} w_{\tilde{\tau}^*} g_{\tilde{\tau}^*} d\tilde{\tau}^* . \tag{3.2}$$

Figure 10 illustrates this capability. In this example, the model is presented with two alternatives that predict a valuable outcome but with different magnitudes and different time courses. When the decision maker approaches the choice with a narrow temporal window, as in the case where the train will arrive in 15 minutes, choice A is more valuable. However, when choosing using a broader temporal window, as in the case where the train will arrive in one hour, choice B is more valuable.

A temporal representation of the future enables not only decision making with different temporal horizons, but also decision making based on relatively complex temporal demands. Consider the case where an outcome is not valuable in the immediate future, but becomes valuable after some time has passed; for instance, perhaps one is not hungry now but will be hungry in an hour. These capabilities are comparable to those offered by model-based RL. However, as discussed above, the representation of the future is scale invariant and can be computed rapidly.

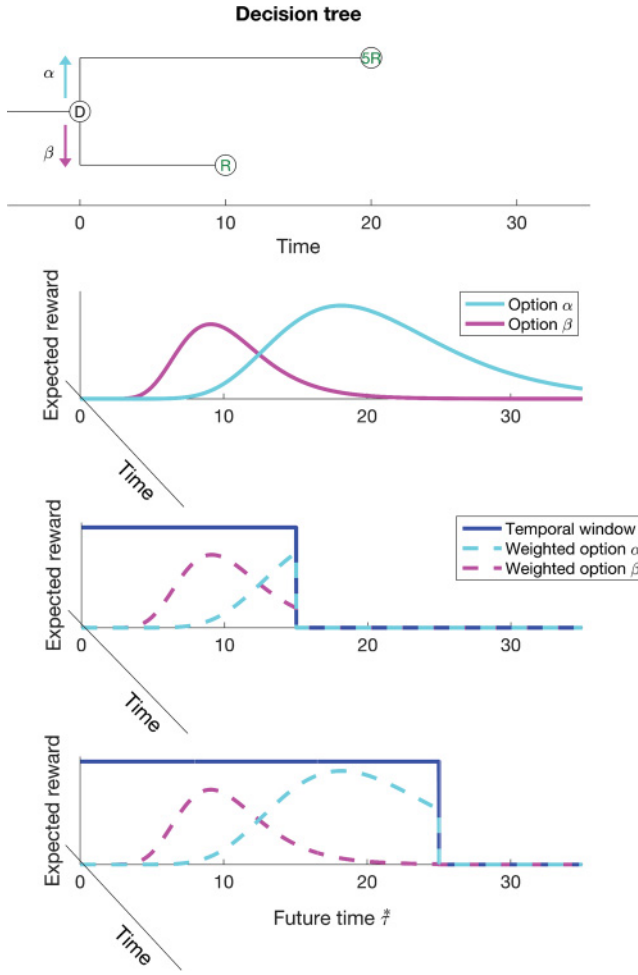

Figure 10: Temporally flexible decision making. Consider an agent faced with two options, α and β, that differ in the time course over which they predict reward (top). Note that β (magenta) predicts a larger reward, but further in the future, than does α (cyan). Representing a function over future time enables the agent to make decisions that incorporate information about the value of the outcome to the agent as a function of time. For instance, under some circumstances, an agent might have more or less time to exploit a food source before some other pressing engagement. The bottom two panels illustrate the value computation with each of two temporal windows. In the middle panel, the temporal window over which the agent can exploit the reward is narrow, and the agent chooses α. In the bottom panel, the temporal window extends further into the future, and the agent chooses β.

4 Behavioral and Neural Predictions

Previous sections presented a method for constructing a compressed estimate of the future. Because this approach is novel, there are no empirical data to definitively evaluate key predictions of this approach. In this section, we describe neural and behavioral predictions and describe how those could be tested experimentally. We also point to recent empirical results, both behavioral and neural, that support the proposed hypothesis, albeit obliquely.

4.1 Cognitive Scanning of the Future. This letter proposes a neural mechanism for constructing a compressed representation of the future. In the cognitive psychology of working memory, prior findings from the short-term judgment of recency (JOR) task suggest that people can scan a compressed representation of the past. For instance, Hacker (1980) presented participants a series of letters rapidly and asked them to evaluate which of two probes was experienced more recently. The critical finding was that the time to choose a probe depended on how far in the past that probe was presented and did not depend on the recency of the other probe. These findings suggested that participants sequentially examine a temporally organized representation of the past and terminate the search when they find a target (see also Muter, 1979; Hockley, 1984; McElree & Dosher, 1993). Furthermore, the time to choose a probe grew sublinearly with how far in the past the probe item was, suggesting that the temporally organized memory representation is compressed (the results were consistent with the hypothesis discussed here that the compression is logarithmic). These findings from the memory literature suggest that an analogous procedure could be used to query participants' expectations about the future. By setting the temporal windowing functions (see equation 3.2) to direct attention to sequentially more distant points in the future, one could sequentially examine an ordered representation of the future.

In order to evaluate whether human participants can scan across a compressed temporally ordered representation of the future, Singh and Howard (2017) trained participants on a probabilistic sequence of letters. After training, the sequence was occasionally interrupted with probes consisting of two letters. Participants were instructed to select the probe that is more likely to appear sooner. If the participants sequentially scan a log-compressed timeline of future events, then this predicts a pattern of results analogous to the findings from the JOR task. Specifically, evidence for sequential scanning would be that the response time in correct trials depends only on the lag of the more imminent probe. Furthermore, in trials in which participants make an error, the response time should depend on the lag of the less imminent probe (this is because if participants have missed the more imminent probe during the scanning process, they will continue scanning until they reach the less imminent probe). Evidence for compression

of the temporally ordered representation of the future would be a sublinear growth of response time with the lag to the probe that is selected. These predictions were confirmed (see Figure 2b; Singh & Howard, 2017).

4.2 Neural Signature of the Compressed Timeline of Future Events. As discussed in section 1, there is ample evidence that neurons in the mammalian brain can be used to decode what happened when in the past (e.g., see Figure 5a; Bolkan et al., 2017; MacDonald et al., 2011; Tiganj et al., 2018; Salz et al., 2016). By analogy, our model predicts that it should be possible to measure neurons that predict what *will* happen when in the future. Because predictions of the future cannot be dissociated from the past, it is possible to have the same future predicted by distinct past events. Consider a situation in which participants are trained on two distinct sequences A, B, C and X, Y, C, and we record after training from a region of the brain representing the future as described by equation 2.13. The model predicts that a common population of neurons (coding for C two steps in the future) should activate when either A or X is presented. The response to the probe stimuli prior to learning of the sequences serves as a control. Similarly, a distinct population (coding for C one step in the future) will be activated when either B or Y is presented. In analogy to sequences of firing triggered by past events (Tiganj et al., 2018), this outcome would imply that similar sequences of neural firing anticipate similar outcomes (see Figure 5b).

5 Discussion

In this letter, we show that given a compressed representation of the past, a simple associative mechanism is sufficient to enable one to generate a compressed representation of the future. A compressed representation of the past has been extensively observed in the brain in many brain regions (MacDonald et al., 2011; Jin, Fujii, & Graybiel, 2009; Tiganj et al., 2018). The associative mechanism we use can be understood as simple Hebbian association. The representation that is generated by this method has many potentially desirable computational properties.

Because the representations of the past and the future are both scale-invariant, it is not necessary to have a strong prior belief about the relevant timescale of the problem one is trying to solve. A scale-invariant learning agent ought to be able to solve problems in a wide range of learning environments. While it remains to be shown that the form of compression of temporal sequences in the brain is quantitatively scale invariant (rather than merely compressed), scale invariance is a design goal that can be implemented in artificial systems.

Because the method directly estimates a function over future states, rather than an integral over future states, decision makers can make adaptive decisions that take into account the time of future outcomes. The future timeline constructed using this method differs from traditional

model-based approaches. After the association matrix **M** has been learned, the computation of the future trajectory is computationally efficient and can be accomplished in one parallel operation. **M** can be learned rapidly, allowing even one-shot learning, unlike, for instance, approaches based on backpropagation. In addition, the logarithmic form for the future means that even if the decision maker queries the representation sequentially, the amount of time to access a future event goes up sublinearly. Recent behavioral evidence from human subjects shows just this result (Singh & Howard, 2017).

Because the method treats time as a continuous variable, there is no need to discretize time. That is, the "distance" between two states need not be filled with other states. In TD learning, error propogates backward from one state to a preceding state via a gradient along intervening states. Using the method in this letter, one can learn that A predicts B separated by, say, 17.4 s without having to define a set of discrete states that intervene. The number of presentations necessary to establish a relationship between two stimuli in **M** depends only on their number of pairings but does not depend on the temporal lag.

5.1 Relationship to the Successor Representation. The idea of efficiently computing compressed summaries of the future arises in another approach to RL, based on the successor representation (SR; Dayan, 1993). Instead of estimating cached values (as in model-free approaches) or transition functions (as in model-based approaches), the SR estimates the discounted expected future occupancy of each state from every other state. The SR can then be combined with an estimated reward function to produce value estimates. Thus, this approach permits the computation of values without expensive tree search or dynamic programming, but retains some of the flexibility of model-based approaches by factoring the value function into predictive and reward components. From a neurobiological and psychological point of view, several lines of evidence have suggested that the brain might use such a representation to solve RL problems (Momennejad et al., 2017; Stachenfeld, Botvinick, & Gershman, 2016).

The SR has many interesting computational properties, but it still runs afoul of the issues raised in this letter. In particular, the SR assumes exponential discounting and consequently imposes a timescale. If the world obeys a Markov process at the assumed timescale, then the SR will be able to efficiently solve RL problems. However, as we pointed out, realistic environments consist of problems occurring at many different scales. Moreover, effective decision making requires explicit information about the time at which stimuli are expected to occur. Thus, effective RL in the real world may require more temporal flexibility than what the SR can provide.

5.2 Relationship to Models of Episodic Memory and Planning. RL models have long utilized a rich interplay between planning, action

selection, and prediction of future outcomes (e.g., Sutton & Barto, 1998). Gershman and Daw (2017), building on an earlier proposal by Lengyel and Dayan (2008), proposed that retrieval of specific instances from memory could enhance RL-based decision making. In psychology, the ability to consciously retrieve specific instances from one's life is referred to as episodic memory (Tulving, 1983). Episodic memory could enhance the capabilities of RL-based models by enabling single-trial learning and bridging across multiple experiences with the same stimulus to discover relationships among temporally remote stimuli (Bunsey & Eichenbaum, 1996; Cole, Barnet, & Miller, 1995; Wimmer & Shohamy, 2012).

Episodic memory has also been proposed to share a neural substrate with what is referred to as episodic future thinking (Tulving, 1985; Schacter, Addis, & Buckner, 2007). Recovery of an episodic memory results in vivid recall of one's past self in a particular spatiotemporal context different from one's present circumstances. Episodic future thinking is defined as imagination of one's future self in a circumstances different from the present. Notably, behavioral and neuroimaging work shows that amnesia patients who are impaired at episodic memory also show deficits in episodic future thinking and that the brain regions engaged by episodic memory performance overlap with the regions engaged by episodic future thinking (Addis, Wong, & Schacter, 2007; Hassabis, Kumaran, Vann, & Maguire, 2007; Palombo, Keane, & Verfaellie, 2015).

Our approach suggests the first steps toward a computational bridge between episodic memory for the past and planning based on future time. In this letter, we showed that a temporal history $\tilde{\mathbf{f}}$ can be used to generate a prediction of the future via an associative memory. The sequentially activated neurons predicted by $\tilde{\mathbf{f}}$ strongly resemble sequentially activated "time cells" measured in the hippocampus (MacDonald et al., 2011; Pastalkova et al., 2008), a brain region implicated in episodic memory. Moreover, the present approach is closely related to the temporal context model, a computational approach that has been applied to behavioral results in a range of episodic memory paradigms (TCM; Howard & Kahana, 2002; Sederberg, Howard, & Kahana, 2008; Polyn, Norman, & Kahana, 2009; Gershman, Moore, Todd, Norman, & Sederberg, 2012). In TCM, items are bound to the prevailing temporal context present when the item appeared via an associative context-to-item matrix. The temporal history $\tilde{\mathbf{f}}$ plays a role very similar to temporal context in TCM, although in TCM, temporal context is an exponentially weighted sum over recent experience that introduces a scale rather than the scale-invariant representation of the past $\tilde{\mathbf{f}}$.

The major departure of our model from TCM is that we have not enabled recovery of a previous history by an item and used to cue future outcomes. That is, one might imagine a model in which, rather than cueing $\mathbf{M}$ with a particular state α, one enables state α to recover a previous state of $\tilde{\mathbf{f}}$ that preceded α and then use that recovered temporal history to predict future

outcomes. This kind of mechanism not only enables TCM to account for the contiguity effect in episodic memory but also allows flexible learning across similar events (Howard, Fotedar, Datey, & Hasselmo, 2005). Future work should explore to what extent a similar contextual reinstatement process, instead in this case reinstating the compressed scale-free representation of the past (Howard et al., 2015), would help speed up learning or transfer of knowledge and predictions as an agent explores a novel world in similar, but not identical, trajectories (Gershman, 2017).

Acknowledgments

We gratefully acknowledge discussions with Karthik Shankar and Ida Momennejad. This work was supported by NIBIB R01EB022864, NIMH R01MH112169, NIH R01- 1207833, MURI N00014-16-1-2832, and NSF IIS 1631460.

References

Addis, D. R., Wong, A. T., & Schacter, D. L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, *45*(7), 1363–1377.

Adler, A., Katabi, S., Finkes, I., Israel, Z., Prut, Y., & Bergman, H. (2012). Temporal convergence of dynamic cell assemblies in the striato-pallidal network. *Journal of Neuroscience*, *32*(7), 2473–84. doi:10.1523/JNEUROSCI.4830-11.2012

Akhlaghpour, H., Wiskerke, J., Choi, J. Y., Taliaferro, J. P., Au, J., & Witten, I. (2016). Dissociated sequential activity and stimulus encoding in the dorsomedial striatum during spatial working memory. *eLife*, *5*, e19507.

Balsam, P. D., & Gallistel, C. R. (2009). Temporal maps and informativeness in associative learning. *Trends in Neuroscience*, *32*(2), 73–78.

Beck, J. M., Latham, P. E., & Pouget, A. (2011). Marginalization in neural circuits with divisive normalization. *Journal of Neuroscience*, *31*(43), 15310–15319.

Bliss, T. V., & Collingridge, G. L. (1993). A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature*, *361*(6407), 31.

Bolkan, S. S., Stujenske, J. M., Parnaudeau, S., Spellman, T. J., Rauffenbart, C., Abbas, A. I., . . . Kellendonk, C. (2017). Thalamic projections sustain prefrontal activity during working memory maintenance. *Nature Neuroscience*, *20*(7), 987–996.

Brown, S., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539.

Bunsey, M., & Eichenbaum, H. B. (1996). Conservation of hippocampal memory function in rats and humans. *Nature*, *379*(6562), 255–257.

Cole, R. P., Barnet, R. C., & Miller, R. R. (1995). Temporal encoding in trace conditioning. *Animal Learning and Behavior*, *23*(2), 144–153.

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655). doi:10.1098/rstb.2013.0478

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

Gershman, S. J. (2017). Predicting the past, remembering the future. *Current Opinion in Behavioral Sciences*, *17*, 7–13.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*, 101–128.

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.

Green, L., & Myerson, J. (1996). Exponential versus hyperbolic discounting of delayed outcomes: Risk and waiting time. *American Zoologist*, *36*(4), 496–505.

Green, L., & Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, *130*(5), 769.

Hacker, M. J. (1980). Speed and accuracy of recency judgments for events in short-term memory. *Journal of Experimental Psychology: Human Learning and Memory*, *15*, 846–858.

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences USA*, *104*(5), 1726–1731. doi:10.1073/pnas.0610561104

Hayden, B. Y. (2016). Time discounting and time preference in animals: A critical review. *Psychonomic Bulletin and Review*, *23*(1), 39–53.

Hockley, W. E. (1984). Analysis of response time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 598–615.

Howard, M. W., Fotedar, M. S., Datey, A. V., & Hasselmo, M. E. (2005). The temporal context model in spatial navigation and relational learning: Toward a common explanation of medial temporal lobe function across domains. *Psychological Review*, *112*(1), 75–116.

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

Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., & Eichenbaum, H. (2014). A unified mathematical framework for coding time, space, and sequences in the hippocampal region. *Journal of Neuroscience*, *34*(13), 4692–4707. doi:10.1523/JNEUROSCI.5808-12.2014

Howard, M. W., & Shankar, K. H. (2018). Neural scaling laws for an uncertain world. *Psychological Review*, *125*, 47–58. doi:10.1037/rev0000081

Howard, M. W., Shankar, K. H., Aue, W., & Criss, A. H. (2015). A distributed representation of internal time. *Psychological Review*, *122*(1), 24–53.

Jin, D. Z., Fujii, N., & Graybiel, A. M. (2009). Neural representation of time in cortico-basal ganglia circuits. *Proceedings of the National Academy of Sciences*, *106*(45), 19156–19161.

Kurth-Nelson, Z., & Redish, A. D. (2009). Temporal-difference reinforcement learning with distributed representations. *PLoS One*, *4*(10), e7362.

Lengyel, M., & Dayan, P. (2008). Hippocampal contributions to control: The third way. In *Advances in neural information processing systems* (pp. 889–896). Cambridge, MA: MIT Press.

Lisman, J., Schulman, H., & Cline, H. (2002). The molecular basis of CAMKII function in synaptic and behavioural memory. *Nature Reviews Neuroscience*, *3*(3), 175–190.

Liu, Y., Tiganj, Z., Hasselmo, M. E., & Howard, M. W. (2018). A neural microcircuit model for a scalable scale-invariant representation of time. *Hippocampus*, *2018*, 1–15. https://dx.doi.org/10.1002/hipo.22994

Ludvig, E. A., Sutton, R. S., & Kehoe, E. J. (2008). Stimulus representation and the timing of reward-prediction errors in models of dopamine system. *Neural Computation*, *20*, 3034–3054.

Ludvig, E. A., Sutton, R. S., & Kehoe, E. J. (2012). Evaluating the TD model of classical conditioning. *Learning and Behavior*, *40*(3), 305–319.

MacDonald, C. J., Carrow, S., Place, R., & Eichenbaum, H. (2013). Distinct hippocampal time cell sequences represent odor memories immobilized rats. *Journal of Neuroscience*, *33*(36), 14607–14616.

MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal "time cells" bridge the gap in memory for discontiguous events. *Neuron*, *71*(4), 737–749.

Mau, W., Sullivan, D. W., Kinsky, N. R., Hasselmo, M. E., Howard, M. W., & Eichenbaum, H. (2018). The same hippocampal CA1 population simultaneously codes temporal information over multiple timescales. *Current Biology*, *28*(10), 1499–1508.

McElree, B., & Dosher, B. A. (1993). Serial recovery processes in the recovery of order information. *Journal of Experimental Psychology: General*, *122*, 291–315.

Mello, G. B. M., Soares, S., & Paton, J. J. (2015). A scalable population code for time in the striatum. *Current Biology*, *25*(9), 1113–1122.

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

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680.

Muter, P. (1979). Response latencies in discriminations of recency. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 160–169.

Palombo, D. J., Keane, M. M., & Verfaellie, M. (2015). The medial temporal lobes are critical for reward-based decision making under conditions that promote episodic future thinking. *Hippocampus*, *25*(3), 345–353.

Pastalkova, E., Itskov, V., Amarasingham, A., & Buzsaki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science*, *321*(5894), 1322–1327.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*, 129–156.

Post, E. (1930). Generalized differentiation. *Transactions of the American Mathematical Society*, *32*, 723–781.

Salz, D. M., Tiganj, Z., Khasnabish, S., Kohley, A., Sheehan, D., Howard, M. W., & Eichenbaum, H. (2016). Time cells in hippocampal area CA3. *Journal of Neuroscience*, *36*(28), 7476–7484.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews, Neuroscience*, *8*(9), 657–661.

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

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*, 893–912.

Shankar, K. H., & Howard, M. W. (2012). A scale-invariant internal representation of time. *Neural Computation*, *24*(1), 134–193.

Shankar, K. H., & Howard, M. W. (2013). Optimally fuzzy temporal memory. *Journal of Machine Learning Research*, *14*, 3753–3780.

Singh, I., & Howard, M. W. (2017). *Scanning along a compressed timeline of the future*. bioRxiv 229617.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2016). *The hippocampus as a predictive map*. bioRxiv, 097170.

Sutton, R. S. (1995). TD models: Modeling the world at a mixture of time scales. In *Proceedings of the International Conference on Machine Learning* (vol. 12, pp. 531–539). Amsterdam: Elsevier.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (vol. 1). Cambridge, MA: MIT Press.

Terada, S., Sakurai, Y., Nakahara, H., & Fujisawa, S. (2017). Temporal and rate coding for discrete event sequences in the hippocampus. *Neuron*, *94*, 1248–1262.

Tiganj, Z., Cromer, J. A., Roy, J. E., Miller, E. K., & Howard, M. W. (2018). Compressed timeline of recent experience in monkey lateral prefrontal cortex. *Journal of Cognitive Neuroscience*, *30*, 1–16.

Tiganj, Z., Hasselmo, M. E., & Howard, M. W. (2015). A simple biophysically plausible model for long time constants in single neurons. *Hippocampus*, *25*(1), 27–37.

Tiganj, Z., Kim, J., Jung, M. W., & Howard, M. W. (2016). Sequential firing codes for time in rodent mPFC. *Cerebral Cortex*, *27*, 1–9. doi:1093/cercor/bhw336

Tulving, E. (1983). *Elements of Episodic Memory*. New York: Oxford University Press.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*(1), 1–12.

Wimmer, G. E., & Shohamy, D. (2012). Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science*, *338*(6104), 270–273. doi:10.1126/science.1223252

Received February 17, 2018; accepted November 22, 2018.