Trends in Cognitive Sciences

CellPress
REVIEWS

Special Issue: Time in the Brain

Review

Integrating Models of Interval Timing and Reinforcement Learning

Elijah A. Petter,[1] Samuel J. Gershman,[2] and Warren H. Meck[1,*]

We present an integrated view of interval timing and reinforcement learning (RL) in the brain. The computational goal of RL is to maximize future rewards, and this depends crucially on a representation of time. Different RL systems in the brain process time in distinct ways. A model-based system learns 'what happens when', employing this internal model to generate action plans, while a model-free system learns to predict reward directly from a set of temporal basis functions. We describe how these systems are subserved by a computational division of labor between several brain regions, with a focus on the basal ganglia and the hippocampus, as well as how these regions are influenced by the neuromodulator dopamine.

Highlights

The relationship between reinforcement learning and interval timing can be followed across Marr's three levels.

At the computational level, interval timing and reinforcement learning are solving the same prediction problem.

At the algorithmic level, both interval timing and reinforcement learning use model-based and model-free learning.

At the implementation level, interval timing and reinforcement learning use striatal, hippocampal, and cortical networks to learn about and predict the state of the environment.

Reinforcement Learning Depends on Temporal Processing

To predict and maximize future reward, an agent must have a sense of time. Many tasks require precisely timed actions, generalization of reward predictions across different intervals, or an accurate representation of the temporal order of events. Consider, for example, the problem of learning to swing a baseball bat. The batter must coordinate a temporally precise sequence of motor commands, but this sequence must be flexible enough so that it can adapt to different pitchers without learning from scratch. The batter can also take advantage of his/her expectations about what happens when, by making anticipatory adjustments to his/her posture as he/she collects information about the ball's trajectory. To be effective, **interval timing** (**IT**; see Glossary) mechanisms have to be flexible in terms of the range of durations that can be timed, and how quickly new temporal criteria can be incorporated into decision-making processes [1,2].

Computational theories of **reinforcement learning (RL)** have begun to incorporate a precise sense of time by building a bridge with theories of temporal processing [3–6]. Guided by Marr's three levels of analysis [7,8] this review will discuss the integration of RL with temporal processing at computational, algorithmic, and neural levels of analysis. We first outline the computational goal of RL, and how this relies critically on IT, particularly in the seconds-to-minutes range. We then discuss how different algorithmic solutions to the RL problem use temporal information. Finally, we relate these algorithms to their underlying neural systems, showing how consideration of temporal processing can resolve a number of limitations and lacunae in our understanding of RL.

The Problem of Time in Reinforcement Learning

The goal of RL is to maximize expected cumulative future reward (i.e., value). This requires the agent to solve a 'prediction problem', which requires estimating the value of each possible action, and an 'optimization problem', which involves balancing exploration and exploitation to select the optimal action. More formally, the prediction problem is to compute the value of each state–action pair, where the **state** distills those aspects of the environment necessary for

[1]Department of Psychology and Neuroscience, Duke University, Durham, NC, USA
[2]Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA, USA

*Correspondence:
whmeck@duke.edu (W.H. Meck).

CrossMark

predicting rewards. A classical assumption is that the time evolution of states and rewards can be described by a **Markov decision process (MDP)**, according to which state transitions and reward depend only on the current state (Box 1). This Markov assumption enables the use of efficient RL algorithms, as we discuss in the following section.

The Markov assumption also makes explicit the requirements for temporal representation. All temporal dynamics must be captured by the state-transition function, which means that the state representation must encode the time-invariant structure of the environment. Consider, for example, a simple Pavlovian conditioning protocol in which a conditioned stimulus (CS; e.g., tone) is followed by a fixed interstimulus interval (ISI) that terminates with the delivery of an unconditioned stimulus (US; e.g., food). The US is then followed by a fixed intertrial interval (ITI). In discrete time, the temporal dynamics can be approximated by breaking the intervals into short 'substates' that are activated sequentially after stimulus onset (see [3] for an alternative continuous time approach). This framework can be generalized to random ISIs and ITIs by assigning transition probabilities between substates such that the Markov process approximates the corresponding interval distribution. The substates encode a time-invariant structure in the sense that the state transition function does not change as the agent proceeds through the task; only the activated substate changes over time.

This example illustrates how an agent needs to anticipate when the reward will be delivered to solve the RL problem. In this case, the agent must explicitly encode time intervals as part of its state representation. Thus, IT is an integral part of RL, cutting across different algorithmic solutions. By contrast, different algorithms make use of time representation in different ways. **Model-free** algorithms use the time representation as a basis set for approximating the value function. Most commonly, this means approximating values as linear combinations of basis functions defined over time. In some instances, however, nonlinear function approximators, such as **recurrent neural networks** [9], may be more effective and biologically realistic representations of time. **Model-based** algorithms use the time representation to compute values by simulating the environmental dynamics. This is more computationally intensive than model-free algorithms, but endows model-based algorithms with more flexibility, because parameter changes in the internal model of temporal structure will immediately change the value estimates. Yet another class of algorithms learns a 'predictive map' that encodes long-range temporal relationships between states [10–13]. This can be combined with reward expectations to form a semiflexible value estimate. For the sake of brevity, however, we will focus on IT in model-based and model-free algorithms.

Timing in Model-Free Reinforcement Learning

Model-free algorithms directly estimate the value function by interacting with the environment. The canonical example is temporal difference (TD) learning [14], which uses the discrepancy between received and expected reward (the TD error) to update its value estimates (see Box 1 for technical details). This algorithm has been influential in neuroscience due to the correspondence between the TD error and the firing of midbrain dopamine neurons [15,16]. The value function is thought to be encoded at corticostriatal synapses, where plasticity is modulated by dopamine release similar to the circuits subserving IT [17–21]. This dopamine-dependent plasticity functions within a precise time window (i.e., 0.3–2.0 s), which has been demonstrated using optical activation of dopaminergic and glutamatergic afferents [22].

TD learning can be generically applied to any state representation; the first generation of models [15,23] adopted the substate representation described above, also known as 'complete serial compound'. However, later work highlighted a number of empirical problems with this

Glossary

Belief state: the posterior distribution over hidden states conditional on sensory data.

Bias–variance trade-off: bias is the degree to which a model systematically underfits the training data; variance is the degree to which a model overfits noise in the data. The two forms of error trade-off against each other, such that reducing bias can lead to greater variance.

Clock speed: the speed at which internal time progresses. Changes in clock speed result in altered temporal perception. Faster clock speed results in earlier temporal reproduction, whereas slower clock speeds lead to later temporal reproduction.

Interval timing: the study of temporal processing and the scaling of durations from any sensory modality in the range of hundreds of milliseconds to minutes. See [85] for additional details.

Markov decision process (MDP): a mathematical framework for describing environments in reinforcement learning. An MDP consists of states (s), actions (a), a reward function that probabilistically maps state–action pairs to reward, and a state transition function that probabilistically maps state–action pairs to the next state. This definition implies that rewards and transitions are conditionally independent of the agent's history given the current state and action [86].

Model-based learning: a class of algorithms that learn an internal model of the environment, consisting of a reward function and a state transition function.

Model-free learning: a class of algorithms that learn reward predictions from direct interactions with the environment.

Peak-interval procedure: the peak-interval procedure is a discrete-trial duration re-production procedure. A percentage of trials are fixed-interval trials, where the subject's first response after a target duration (e.g., 30 s) is reinforced. There are also unreinforced probe trials interleaved with the fixed-interval trials. The omission of reward allows timing behavior to be assessed without the interruption of representation [3,4,24,25]. From a computational perspective, the discrete and bounded nature of substates is disadvantageous because it does not allow values learned for one substate to gracefully generalize across time except by the slow transfer of value between adjacent substates that is built into the TD algorithm. Intuitively, information about the value of one substate should be informative about the value of an adjacent substate, even in the absence of new learning.

This issue highlights the tension between representational constraints and learning constraints. Specifically, an agent requires a state representation that can represent the correct value function, but it is necessary to restrict the function class when an agent needs to generalize from limited experience. The **bias–variance trade-off** [26,27] provides a formal lens through which to view this issue: an RL agent can reduce bias (the difference between the learned and true value function) by finely decomposing the state space, but only at the expense of increasing variance (sensitivity to noise in the training data). Balancing bias and variance requires state representations that permit some degree of generalization. In the context of timing, this balance can be achieved using state representations that change smoothly over time. Temporal smoothness acts as a soft constraint on the function class, effectively low-pass filtering the value function and thereby suppressing noise.

We can think about the bias–variance trade-off in neurobiological terms by imagining a set of neurons whose firing represents the activation of particular substates. If the neurons are narrowly tuned, and only respond during the interval corresponding to their preferred substate, then bias will be low and variance will be high: the value function can be precisely represented at each time interval, but the estimated values will be unreliable. If the neurons are broadly tuned, and thus respond partially to neighboring time intervals, then bias will increase and variance will be reduced: the value function can only be imprecisely represented at each time interval, but the estimated values will be more reliable because the neurons collect partial credit for rewards received outside their preferred time interval, effectively increasing the amount of data available to each neuron.

This logic can be extended one step further by considering the fact that most tasks only need accurate IT over a particular range of durations. Moreover, many studies have found that timing noise increases with interval duration in a proportional manner (scalar property – e.g., [28]), which means that there is a limit on the accuracy with which value functions can be represented for very long durations. Thus, neurons tuned to longer durations are not useful for predicting future reward. This suggests that the temporal precision of the state representation should be concentrated around shorter durations. One implementation of this idea uses ‘microstimuli’ (Figure 1) having Gaussian temporal receptive fields whose width increases and amplitude decreases with the preferred duration [24,25; see also 29]. Recent experiments have presented evidence for microstimulus-like temporal receptive fields, also known as ‘**time cells**’, in several areas, including the striatum [30,31], hippocampus [32–35], and prefrontal cortex [36]. These neural representations can in some cases quantitatively predict variation in timing behavior [37].

A striking feature of ‘time cells’ noted in several reports [33,30,38] is their ability to rescale when the intervals are changed dramatically. For example, striatal neurons encode a range of intervals when a subject is trained on a fixed-interval schedule, and these same neurons expand the range of intervals when the fixed interval is increased. Further, this rescaling of neurons has been observed on a trial-by-trial basis in both the caudate and the medial frontal cortex of a monkey performing a two-interval reproduction task [39]. This ability to display time-invariant neural activity for multiple durations can be modeled using a recurrent neural network

reward, allowing for the individual trials analysis of ‘stop’ and ‘start’ responses [87].

Recurrent neural network: a network in which neurons are connected sequentially, giving rise to temporal dynamics.

Reinforcement learning: the problem of learning to maximize cumulative future rewards [88].

Reward prediction error: the discrepancy between received and expected reward. This error is often attributed to the firing patterns of midbrain dopamine neurons (e.g., [89]). Recent findings, however, have challenged this long-held view by attributing the role of striatal dopamine to movement kinematics (e.g., head velocity) in unrestrained subjects performing in behavioral tasks involving the delivery of reward [90].

Scalar-timing noise: variance of responses increases in proportion to the duration that is being timed. Time cells also demonstrate scalar noise, with larger temporal receptive fields for longer durations.

State: the information about the current environment sufficient to predict future reward.

Structure learning: the problem of discovering the state space underlying sensory data.

Temporal bisection procedure: this is a duration classification task in which subjects are presented with either a ‘short’ or a ‘long’ signal duration after which a response (e.g., left button) will be reinforced if the ‘short duration’ had been presented, whereas a different response (e.g., right button) will be reinforced if the ‘long’ duration had been presented. Once subjects acquire this discrimination, they are presented with a range of intermediate durations distributed between the ‘short’ and ‘long’ durations. Analysis is generally performed on the normalized slope of the function (i.e., Weber fraction), and the duration at which subjects have a 50% probability of responding ‘long’ (i.e., the point of subjective equality).

Time cells: neurons that peak in activation at a characteristic delay from a salient event. These have been found in the striatum, hippocampus, and cortical regions, among others.

Box 1. Markov Decision Processes

An MDP is defined by reward function *R*(s,a) mapping state–action pairs to scalar rewards, a state transition function *T* (s′|s,a) mapping state–action pairs to a probability distribution over the next state, and a discount factor γ specifying the agent's time preference (weighing proximal rewards greater than distal rewards; Figure I). The Markovian (history-independent) structure allows the value function to be expressed in an algorithmically convenient form known as the Bellman equation [88]:

$$Q_\pi(s_t,a_t) = R(s_t,a_t) + \gamma \sum_{s_{t+1}} T(s_{t+1}|s_t,a_t) Q_\pi\left(s_{t+1,}\pi(s_{t+1})\right)$$

where *t* indexes time, *s* denotes the state, *a* denotes the action, and π is a policy mapping from states to actions (taken here to be deterministic for simplicity).

Model-based algorithms use the reward and state transition functions to iteratively compute an estimate $\hat{Q}_\pi$ of the value function. Intuitively, the Bellman equation specifies a self-consistency relation between values of adjacent states, so we can use the discrepancy between these (estimated) values (i.e., the difference between the right- and left-hand sides of the Bellman equation) to adjust them in the right direction:

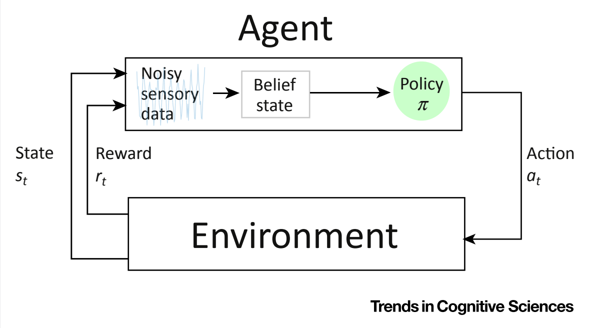
$$\Delta\hat{Q}_\pi(s_t,a_t) \propto R(s_t,a_t) + \gamma \sum_{s_{t+1}} T(s_{t+1}|s_t,a_t) \hat{Q}_\pi\left(s_{t+1,}\pi(s_{t+1})\right) - \hat{Q}_\pi(s_t,a_t)$$

where the proportional symbol indicates that the change in value may be incremental, as determined by a learning rate (not shown here). This equation is an incremental version of the classical value iteration algorithm [88].

In practice, some state transitions are low probability, so it is computationally wasteful to sum over all possible transitions. An alternative is to simulate (take random samples) from the state transition function, thereby constructing a Monte Carlo approximation whose accuracy increases as more samples are simulated. Even drawing a single sample s_{t+1} will lead to an update that eventually converges to the true value function:

$$\Delta\hat{Q}_\pi(s_t,a_t) \propto R(s_t,a_t) + \gamma\hat{Q}_\pi\left(s_{t+1,}\pi(s_{t+1})\right) - \hat{Q}_\pi(s_t,a_t)$$

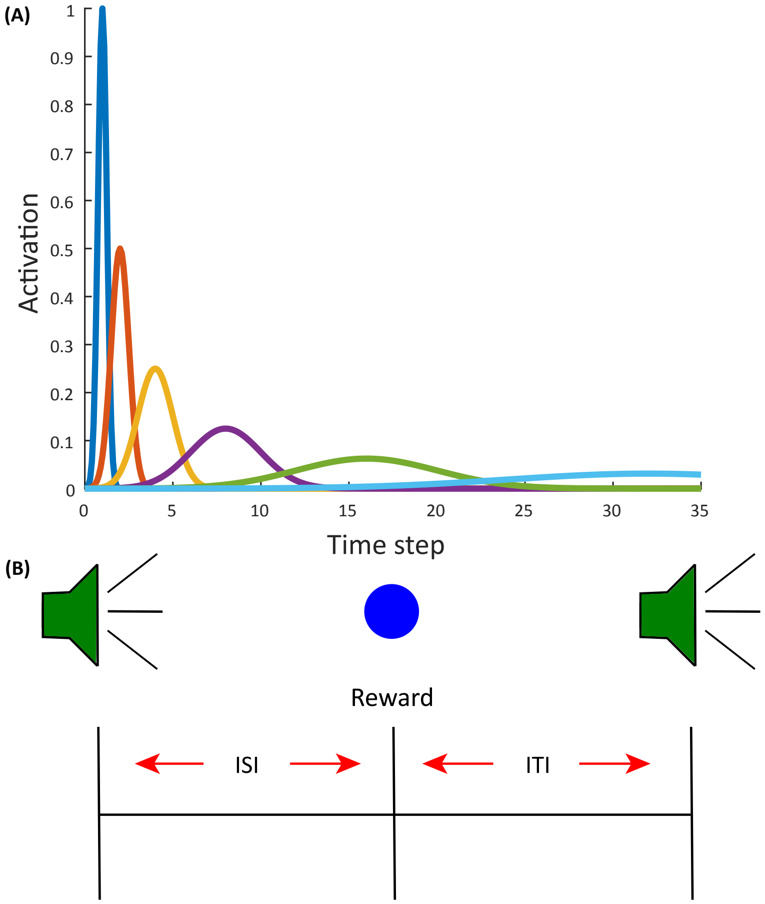
If the agent does not have access to an internal model of the state transition function, this update rule can still be applied, simply by interacting with the environment. That is, an agent can take actions in the environment and experience state transitions, plugging them into the update rule. This corresponds to the TD algorithm [14], the celebrated 'model-free' learning algorithm, where the update is known as the TD error. The intimate relationship between model-based and model-free algorithms via the Bellman equation suggests that they can be productively combined (see [91]). For example, a model-free agent can update value estimates based on both simulated and experienced transitions [4,92].



Trends in Cognitive Sciences

Figure I. The Interaction of Agent and Environment. An agent interacts with the environment through selecting actions, resulting in state transitions, rewards, and sensory evidence, which the agent uses to infer the hidden state. Feedback from the environment is used to update the agent's action selection policy.

that exhibits stereotyped neural trajectories that vary speed based on a gain input (see also [40]). Consequently, it appears as though the transition through neural state space does not change when the temporal criteria are altered, but just the rate at which the trajectory progresses (i.e., **clock speed**).



Trends in Cognitive Sciences

Figure 1. Predicting Reward with Temporal Representations. (A) Learning to anticipate a reward that follows a predictive cue can be accomplished by using a set of 'microstimulus' representations (temporal basis functions; see [24,25]). (B) These basis functions are activated by salient events (e.g., a tone signaling trial start), and peak at a characteristic delay from event onset. They have the benefit of naturally allowing for generalization of value across time. Specifically, the overlapping nature of the basis functions allows partial credit when rewards are received at times of less than maximal activation. Here the width of each basis function scales in proportion to its preferred duration, thus increasing the amount of generalization with longer durations. ISI, interstimulus interval; ITI, intertrial interval.

At least in some cases [33], the emergence of rescaling is gradual, which could reflect slow reinforcement-driven adaptation. The mechanisms underlying this adaptation are still poorly understood. One possibility is that dopamine itself is the teaching signal for adaptation. Extensive pharmacological work has shown that reductions in tonic levels of dopamine cause intervals to be perceived as proportionally longer, in effect causing the IT mechanisms to run faster [41,42]. Further, tonic dopamine changes affect intertemporal choice, where subjects choose between rewards available at different points in time. Dissociations have been shown between IT and intertemporal choice following the administration of dopaminergic and serotonergic drugs. In these experiments, variations in timing (clock speed) were able to mediate intertemporal choice via dopaminergic inputs. By contrast, a separate serotonergic system was able to affect intertemporal choice without affecting IT directly [43].

Recent optogenetic manipulations of dopamine neurons during a **temporal bisection procedure** [44] appear to paint a different picture, with decreases in phasic dopamine activity causing intervals to be perceived as shorter under some conditions. Interestingly, the strongest effects were shown around the point of subjective equality (PSE), or where rats had a 50% chance of classifying the duration as either long or short. The PSE can be thought of as the duration that has the highest level of uncertainty, and as a consequence, might be most affected by increased phasic dopamine activity.

Some RL theories have posited different computational roles for tonic and phasic dopamine, with phasic activity reporting **reward prediction errors** and tonic activity reporting average reward [45], but it remains unclear how this division of labor resolves the discrepancy between pharmacological and optogenetic dopamine manipulations. Another possibility is that tonic dopamine controls the precision (inverse variance) of sensory measurements [46–48]. This could explain why decreasing tonic dopamine shortens subjective duration relative to physical time: when precision is low, the duration estimate is regularized towards the prior (see below for a discussion of central tendency effects), which will be downward when intervals tend to be short. However, this does not explain the effects of phasic dopamine manipulations – an important open problem for experimental and theoretical research.

Timing in Model-Based Reinforcement Learning

In contrast to the relative inflexibility of model-free algorithms, which must incrementally relearn value estimates when the environment changes, model-based algorithms are capable of rapid and dramatic changes in value estimates. This flexibility derives from the fact that changes in the environment (assuming it is modeled as an MDP) can be captured by local modifications of the transition function, in contrast to the value function which encodes long-term dependencies (Box 1).

Studies of IT have provided evidence for such flexibility. For example, subjects trained in a **peak-interval procedure** show rapid acquisition in response to abrupt changes in the criterion duration [49,50]. Similarly, rapid acquisition of timing behavior has been reported in Pavlovian delay conditioning experiments [51].

The rapid nature of this learning is further apparent in experiments that used a hopper-switch task [52–54] in which reward is probabilistically assigned on each trial to one of two hoppers. If reward is assigned to the 'short' hopper, the mouse receives a food pellet by poking the hopper after 3 s have elapsed. If reward is assigned to the 'long' hopper, the mouse receives a pellet by poking the hopper after 9 s have elapsed. The optimal policy depends on the probability distribution over hoppers. Occasionally, this distribution was changed, and the authors estimated that mice took approximately 10 trials to detect the change, after which there was an abrupt shift in their behavior. Thus, temporal information appears to be used to update parameters in an internal model to achieve change detection, rather than the gradual adjustment of behavior anticipated by standard model-free algorithms.

Even more decisive evidence for the flexibility of timing has been provided using a variant of the hopper-switch task in which mice are first trained on each hopper separately (i.e., no switching) and then in the test phase given a choice between hoppers [55]. Despite never having been trained on the two-hopper task, mice almost immediately and stably adopted the optimal policy in the test phase, going first to the 'short' hopper and then, if not rewarded, to the 'long' hopper. Moreover, mice calibrated their switch latency to the probability of reward, staying longer at the 'short' hopper when its reward frequency was higher during the training phase.

Similarly, subjects are capable of spontaneous averaging of intervals. On separate trials rats were trained to associate an auditory stimulus with a short delay to food availability, and a visual stimulus with a long delay. When both modalities were presented simultaneously, subjects averaged the two separate reinforcement times [56].

Internal models of timing can also be exposed by manipulating the distribution of intervals. These manipulations have been performed with an operant conditioning task in which the CS–US interval was drawn from a Gaussian whose mean was held fixed across blocks of trials within a session, but whose variance was varied across blocks [57]. Mice waited longer before approaching the water port when the variance was larger, consistent with the optimal model-based policy. Intuitively, when variance is large, and thus mice are more uncertain, they will wait longer to increase their confidence that the criterion interval has elapsed.

Another source of evidence for model-based learning comes from studies of temporal integration (see [58]; or [59] for a review; see [60] for a modeling approach to some of these phenomenon). In these experiments, subjects are separately trained on different stimulus arrangements, and then are tested on their ability to integrate the temporal relationships between events to form expectations about stimulus arrangements that they have not directly experienced. One example comes from when subjects were initially trained using a backward conditioning arrangement in which a US occurred 5 s before a CS (denoted S1) [61]. The animals then learned about a stimulus arrangement in which S1 was preceded by another CS (S2) such that S2 coterminated with the hypothetical onset of the US, which was omitted. Subjects demonstrated a significant conditioned response to S2 in a final test phase, even though they had never been reinforced following presentation of S2, suggesting that they had formed a temporal representation of the event structure.

In summary, a number of experimental paradigms have shown that subjects learn internal models of timing, which they employ in flexible ways – rapidly adjusting behavior to changes in the internal model, and integrating separately trained fragments of event knowledge. These findings are broadly consistent with computational models of model-based RL (e.g., [62]), but do not strongly favor specific algorithms (though see [63] for a recent attempt at empirically identifying stronger algorithmic assumptions).

State Uncertainty

The algorithms discussed above all assume that the agent has access to the state of the environment. However, state uncertainty is inevitable because we can only make inferences about our environment based on noisy sensory data [64]. In IT, these noisy estimates are apparent from **scalar-timing noise** [28]. State uncertainty can also arise from dynamic environments; for example, if the US is occasionally omitted in a Pavlovian conditioning task with random ISIs (Figure 2), the subject will be unsure whether the US is late (and hence the subject is still in the ISI state) or if it is omitted (and hence the subject has transitioned to the ITI state).

One way to handle state uncertainty in RL is to define value functions over **belief states**, or probability distributions over states (Box 2). The same RL algorithms can be applied to this belief state, which is updated according to Bayes' rule after each new sensory observation [65]. The belief state model leads to the hypothesis that TD errors signaled by dopamine will reflect the evolution of state uncertainty over time [3]. Two recent studies [5,6] have tested this hypothesis in mice (see also [66]).

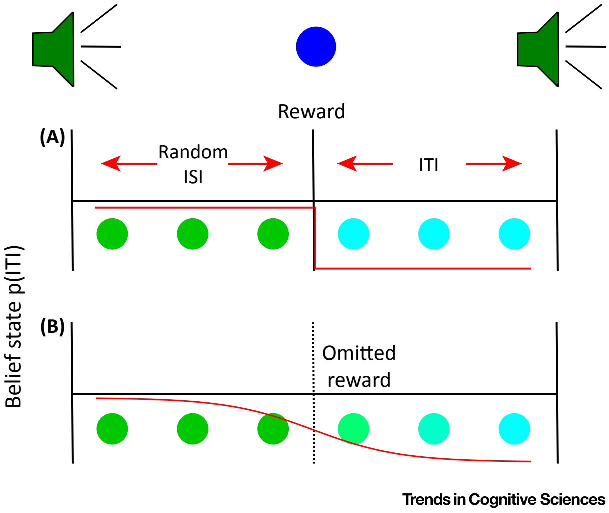


Figure 2. Interval Timing and Belief States. Interactions between interval timing and belief states (probability distributions over states) can be understood through a comparison of Pavlovian tasks with and without state uncertainty. When a reward follows a tone with 100% probability (A), the animal always knows that it is in the interstimulus interval (ISI) state before reward [p(ITI) = 1] and in the intertrial interval (ITI) state after reward [p(ITI) = 0]. As time progresses in the ISI state, the reward becomes more likely, and subjects are less surprised when it is delivered. This corresponds to a monotonic decline in the dopamine reward prediction error signal as a function of ISI (Starkweather *et al*. [5,6]). When rewards are occasionally omitted (B), subjects have uncertainty about whether they are in the ISI or the ITI state. This uncertainty can be captured by a dynamic belief state. As time progresses without a reward, subjects become increasingly confident that the state has silently transitioned to the ITI, and thus their reward expectations will eventually decrease. In this case, reward prediction errors should increase as a function of time, consistent with experimental findings (Starkweather *et al*. [5,6]).

Using a Pavlovian conditioning task with random ISIs, researchers showed that the response of dopamine neurons to reward delivery was sensitive to both the ISI and the probability of reward omission. When the omission probability was 0, and hence the subject had timing uncertainty but not uncertainty about whether it was in the ISI or ITI, dopamine reward responses declined with the ISI. This agrees with previous findings [67] as well as with the TD theory: as time elapses, the subject becomes increasingly confident that the reward will be delivered, and hence the reward is less surprising (lower prediction error). By contrast, when reward was omitted on 10% of the trials the dopamine reward response increased over time. Because of ISI/ITI uncertainty, the TD algorithm operating on belief states predicts this increase. Intuitively, the more time that elapses, the more confident the subject becomes that the hidden state has silently transitioned from the ISI to the ITI.

What are the origins of the belief state inputs to dopamine neurons? Research aimed at addressing this question [6] reversibly inactivated the medial prefrontal cortex (mPFC) while simultaneously recording dopamine neurons. Somewhat surprisingly, inactivation did not alter the temporal modulation of dopamine reward responses in the no-omission, indicating that temporal information from the mPFC was not required for timing in dopamine neurons. By contrast, inactivation in the omission condition largely eliminated the temporal modulation of dopamine responses. Computational modeling using belief state RL indicated that this effect could be understood as an impairment of the belief state representation rather than of IT. In particular, mPFC inactivation appeared to ‘freeze’ the belief state, preventing it from changing in response to reward omission.

Box 2. Reinforcement Learning under State Uncertainty

The algorithms reviewed in Box 1 rely crucially on the agent having access to the state of the environment, but in many real-world scenarios the agent may only receive noisy sensory data about the state. The agent cannot simply learn values for the sensory data, because transitions in the sensory space do not obey the Markov property and hence the Bellman equation is violated. One elegant solution to this problem is to define a belief state, $b_t(s) = P(s_t = s|o_t)$, the posterior probability distribution over hidden states conditional on the sensory data o_t. We can then define a value function over belief states that is Markovian, obeying the following Bellman equation:

$$Q_\pi(b_t, a_t) = R(b_t, a_t) + \gamma \sum_{b_{t+1}} T(b_{t+1}|b_t, a_t) Q_\pi\left(b_{t+1,}\pi(b_{t+1})\right)$$

where the reward and state transition functions are now defined as 'marginals' over the hidden state (see [93] for more details). The algorithms in Box 1 can be applied to this belief state MDP.

This formalism can be applied to various classes of state uncertainty. For example, in reversal learning, the environment consists of two possible hidden states (A and B) distinguished by their reward functions, so the belief state can be represented by a scalar, $b_t = P(s_t = A|o_t)$. In a change detection paradigm, the environment consists of a stochastic sequence of states (A, B, C, D, . . .) that never repeat. In this case, the belief state is in theory infinitely long, but can be efficiently approximated by considering only the probability of a state change on each trial [94], in which case the belief state is again scalar, $b_t = P(s_t \neq s_{t-1}|o_t)$. In Pavlovian conditioning paradigms with reward omissions, the environment consists of two hidden states: the ITI and the ISI. Depending on whether these epochs are modeled in continuous time [3] or broken into discrete temporal substates [5,6], the belief state is thus either scalar, $b_t = P(s_t = ITI|o_t)$, or vector valued $b_t(s) = P(s_t = ITI(s)|o_t)$, where *ITI(s)* denotes substate *s* of the ITI, corresponding to a relatively small duration within the ITI.

Taken together, these findings indicate the importance of considering state uncertainty when studying timing. State uncertainty requires individuals to rely on an internal perception of time that is heavily influenced by belief states. In times of low uncertainty (e.g., significant cues) our perception of time is synchronized to external events. Further, time modulates many downstream computations, such as TD errors and belief states, so it is necessary to consider them jointly to disentangle their neural correlates.

Structure Learning

In many environments, agents might not only be uncertain about the hidden state, but also be uncertain about the state space. This means agents must solve a **structure learning** problem simultaneously with the RL problem of maximizing rewards [68]. The formal framework of belief state RL (Box 2) can be naturally extended to the problem of structure learning.

The aforementioned change detection (hopper-switch) paradigms [53,52,63] can be viewed as a form of structure learning. In these paradigms, the environment never returns to exactly the same state, so the belief state is in theory infinite dimensional. However, real-world environments often have recurring structure, where old states can reoccur in addition to the occurrence of new states. The computational problem facing the brain in these environments is therefore somewhat different from change detection. Instead of detecting whether the current observation is generated by either the same state as the previous observation or a new state, structure learning in recurrent environments needs to consider whether a previously encountered state has reoccurred [4,69].

A simple example of this is the transition from a fixed interval to a peak-interval procedure. Initially, the subject is rewarded 100% of the time for a response occurring after a fixed delay from signal onset. Subjects are then transferred to partial reinforcement schedules, and must learn that in addition to the rewarded state, there is a new state where reward is not available.

A recent work [70] provides another example of structure learning in a temporal reproduction task. Human observers were tasked with reproducing the duration of a visual stimulus by holding down a key for an equivalent duration. Consistent with a large experimental literature (e.g., [71–73]), observers exhibited a central tendency effect: relative to the mean duration, long durations were underestimated, and short durations were overestimated [5]. This effect can be understood in terms of a Bayesian analysis [47,71,72,74], according to which noisy timing signals are combined with a prior over duration to form a posterior (i.e., the belief state). The expected duration under this belief state will be intermediate between the observed timing signal and the mean of the prior.

Further, this research [70] asked whether the central tendency effect reflects a single prior learned across environments, or instead reflects separate priors learned for different environments. To answer this question, they interleaved stimuli from different distributions (e.g., sensory modality), and found that the reproduction bias reflected a single combined prior, even when stimuli for the two distributions were presented in distinct spatial locations. By contrast, when distinct motor responses were required for the two distributions the reproduction bias reflected separate priors, indicating no generalization. These results suggest that motor information exerts a stronger influence on structure learning than sensory information (see also [75]).

Discovering the structure of the environment can provide powerful generalization abilities. For example, subjects trained in the peak-interval procedure exhibit a slower acquisition of the peak time for the first criterion change, and more abrupt acquisitions for subsequent changes ([49] – a form of 'learning to learn' [76]). Once subjects have discovered the underlying structure of temporal criterion changes in the task, they become more adept at adapting to these changes.

Another instance of structure learning is the retention that is present when generalizing a previously learned duration across stimulus modality, location, and task demands. This can be demonstrated when learning a visual duration, and then retaining a degree of performance when the stimulus is switched to the auditory modality. Importantly this does not happen immediately, but takes some time for the new modality to resemble the same interval in a different modality [77–80].

Concluding Remarks and Future Directions

Computational theories of RL and IT have matured independently, and only recently have they begun to intersect [4]. This review has highlighted several ways in which RL algorithms use temporal representations, which go beyond the traditional function approximation with a temporal basis set seen in model-free RL theories. Modern RL embraces a broader range of algorithms, including model-based algorithms that use IT to support flexible and rapid changes in behavior.

Many open questions remain. One concerns our assumption that the environment can be modeled as an MDP. Markov dynamics are algorithmically useful, but they are also restrictive. Several lines of research suggest ways to get around the Markov assumption without entirely abandoning algorithmic efficiency, for example, using episodic memory [69] or a collection of timing elements that support learning at multiple scales [81,82].

Another issue concerns the neural integration of IT and RL mechanisms. Specifically, there are distributed brain regions that support the processing of both IT and RL, yet information is integrated across different algorithmic learning systems to form a perceptual whole [83]. One useful way to better understand the interaction of these two fields is through the use of biologically plausible neural networks. Recent work has shown how biologically plausible recurrent neural networks can be used to implement flexible timing [40,84]. These networks are trained using Hebbian rules [84] or supervised learning [40], rather than RL. However, the

Outstanding Questions

How can theoretical accounts of timing and reward processing be integrated into a more general model?

How do distributed networks involved in RL and IT cooperate and compete to form a unified perception of time?

How can the effects of tonic and phasic dopamine on RL and IT be reconciled into a cohesive story?

As recurrent neural networks are capable of timing behavior without reinforcement learning, how can they be integrated into the framework articulated here?

How does the interaction between IT and RL translate from the laboratory to real-world settings?

insights from these network models may prove useful in constructing more biologically plausible and powerful implementations of RL algorithms (see [9] for a step in this direction). Overall, we encourage a more unified approach to the study of IT and RL going forward. It becomes clear the two disciplines are overlapping as they are traced across Marr's three levels of analysis (see Outstanding Questions).

References

1. Allman, M.J. *et al.* (2014) Properties of the internal clock: first- and second-order principles of subjective time. *Annu. Rev. Psychol.* 65, 743–771
2. Petter, E.A. *et al.* (2016) Interactive roles of the cerebellum and striatum in sub-second and supra-second timing: support for an initiation, continuation, adjustment, and termination (ICAT) model of temporal integration. *Neurosci. Biobehav. Rev.* 71, 739–755
3. Daw, N.D. *et al.* (2006) Representation and timing in theories of the dopamine system. *Neural Comput.* 18, 1637–1677
4. Gershman, S.J. *et al.* (2014) Time representation in reinforcement learning models of the basal ganglia. *Front. Comput. Neurosci.* 7, 194
5. Starkweather, C.K. *et al.* (2017) Dopamine reward prediction errors reflect hidden-state inferences across time. *Nat. Neurosci.* 20, 581–589
6. Starkweather, C.K. *et al.* (2018) The medial prefrontal cortex shapes dopamine reward prediction errors under state uncertainty. *Neuron* 98, 616–629
7. Marr, D. and Poggio, T. (1977) From understanding computation to understanding neural circuitry. *Neurosci. Res. Prog. Bull.* 15, 470–488
8. Niv, Y. and Langdon, A. (2016) Reinforcement learning with Marr. *Curr. Opin. Behav. Sci.* 11, 67–73
9. Song, H.F. *et al.* (2017) Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife* 6, e21492
10. Dayan, P. (1992) The convergence of TD (λ) for general λ. *Mach. Learn.* 8, 341–362
11. Momennejad, I. *et al.* (2017) The successor representation in human reinforcement learning. *Nat. Hum. Behav.* 1, 680
12. Russek, E.M. *et al.* (2017) Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Comput. Biol.* 13, e1005768
13. Stachenfeld, K.L. *et al.* (2017) The hippocampus as a predictive map. *Nat. Neurosci.* 20, 1643
14. Sutton, R.S. (1988) Learning to predict by the method of temporal differences. *Mach. Learn.* 3, 9–44
15. Schultz, W. *et al.* (1997) A neural substrate of prediction and reward. *Science* 275, 1593–1599
16. Glimcher, P.W. (2011) Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* 108, 15647–15654
17. Coull, J.T. *et al.* (2011) Neuroanatomical and neurochemical substrates of timing. *Neuropsychopharmacology* 36, 3–25
18. Meck, W.H. (1988) Internal clock and reward pathways share physiologically similar information-processing stages. In *Quantitative Analyses of Behavior: Biological Determinants of Reinforcement* (Commons, M.L. *et al.*, eds) (Vol. 7), pp. 121–138, Erlbaum
19. Meck, W.H. (2006) Neuroanatomical localization of an internal clock: a functional link between mesolimbic, nigrostriatal, and mesocortical dopaminergic systems. *Brain Res.* 1109, 93–107
20. Merchant, H. *et al.* (2013) Neural basis of the perception and estimation of time. *Annu. Rev. Neurosci.* 36, 313–336
21. Reynolds, J.N.J. and Wickens, J.R. (2002) Dopamine-dependent plasticity of cortico-striatal synapses. *Neural Netw.* 15, 507–521
22. Yagishita, S. *et al.* (2014) A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* 345, 1616–1620
23. Sutton, R.S. and Barto, A.G. (1990) Time derivative models of Pavlovian reinforcement. In *Learning and Computational Neuroscience: Foundations of Adaptive Networks* (Gabriel, M. R. and Moore, J.W., eds), pp. 497–537, MIT Press
24. Ludvig, E.A. *et al.* (2008) Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Comput.* 20, 3034–3054
25. Ludvig, E.A. *et al.* (2012) Evaluating the TD model of classical conditioning. *Learn. Behav.* 40, 305–319
26. Geman, S. *et al.* (1992) Neural networks and the bias/variance dilemma. *Neural Comput.* 4, 1–58
27. Glaze, C.M. *et al.* (2018) A bias-variance trade-off governs individual differences in on-line learning in an unpredictable environment. *Nat. Hum. Behav.* 2, 213–224
28. Gibbon, J. *et al.* (1984) Scalar timing in memory. *Ann. N. Y. Acad. Sci.* 423, 52–77
29. Grossberg, S. and Schmajuk, N.A. (1989) Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Netw.* 2, 79–102
30. Mello, G.B. *et al.* (2015) A scalable population code for time in the striatum. *Curr. Biol.* 25, 1113–1122
31. Lusk, N.A. *et al.* (2016) Cerebellar, hippocampal, and striatal time cells. *Curr. Opin. Behav. Sci.* 8, 186–192
32. Pastalkova, E. *et al.* (2008) Internally generated cell assembly sequences in the rat hippocampus. *Science* 321, 1322–1327
33. MacDonald, C.J. *et al.* (2011) Hippocampal "time cells" bridge the gap in memory for discontiguous events. *Neuron* 71, 737–749
34. MacDonald, C.J. *et al.* (2013) Distinct hippocampal time cell sequences represent odor memories in immobilized rats. *J. Neurosci.* 33, 14607–14616
35. Salz, D.M. *et al.* (2016) Time cells in hippocampal area CA3. *J. Neurosci.* 36, 7476–7484
36. Tiganj, Z. *et al.* (2016) Sequential firing codes for time in rodent medial prefrontal cortex. *Cereb. Cortex* 27, 5663–5671
37. Gouvêa, T.S. *et al.* (2015) Striatal dynamics explain duration judgments. *eLife* 4, e11386
38. Emmons, E.B. *et al.* (2017) Rodent medial frontal control of temporal processing in the dorsomedial striatum. *J. Neurosci.* 37, 8718–8733
39. Wang, J. *et al.* (2018) Flexible timing by temporal scaling of cortical responses. *Nat. Neurosci.* 21, 102
40. Goudar, V. and Buonomano, D.V. (2018) Encoding sensory and motor patterns as time-invariant trajectories in recurrent neural networks. *eLife* 7, e31134
41. Agostino, P.V. and Cheng, R.-K. (2016) Contributions of dopaminergic signaling to timing. *Curr. Opin. Behav. Sci.* 8, 153–160
42. Lake, J.I. and Meck, W.H. (2013) Differential effects of amphetamine and haloperidol on temporal reproduction: dopaminergic regulation of attention and clock speed. *Neuropsychologia* 51, 284–292
43. Heilbronner, S.R. and Meck, W.H. (2014) Dissociations between interval timing and intertemporal choice following administration of fluoxetine, cocaine, or methamphetamine. *Behav. Process.* 101, 123–134
44. Soares, S. *et al.* (2016) Midbrain dopamine neurons control judgment of time. *Science* 354, 1273–1277
45. Niv, Y. *et al.* (2007) Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology* 191, 507–520

46. Friston, K.J. *et al.* (2012) Dopamine, affordance and active inference. *PLoS Comput. Biol.* 8.1, e1002327

47. Gu, B.-M. *et al.* (2016) Bayesian optimization of interval timing and biases in temporal memory as a function of temporal context, feedback, and dopamine levels in young, aged, and Parkinson's disease patients. *Timing Time Percept.* 4, 315–342

48. Mitchell, J.M. *et al.* (2018) Dopamine, time perception, and future time perspective. *Psychopharmacology* 1–11

49. Lejeune, H. (1998) Peak procedure performance in young adult and aged rats: acquisition and adaptation to a changing temporal criterion. *Q. J. Exp. Psychol. B* 51, 193–217

50. MacDonald, C.J. *et al.* (2012) Acquisition of "start" and "stop" response thresholds in peak-interval timing is differentially sensitive to protein synthesis inhibition in the dorsal and ventral striatum. *Front. Integr. Neurosci.* 6, 10

51. Kirkpatrick, K. and Church, R.M. (2000) Independent effects of stimulus and cycle duration in conditioning: the role of timing processes. *Anim. Learn. Behav.* 28, 373–388

52. Kheifets, A. and Gallistel, C.R. (2012) Mice take calculated risks. *Proc. Natl. Acad. Sci. U. S. A.* 109, 8776–8779

53. Balci, F. *et al.* (2009) Risk assessment in man and mouse. *Proc. Natl. Acad. Sci. U. S. A.* 106, 2459–2463

54. Gür, E. and Balci, F. (2017) Mice optimize timed decisions about probabilistic outcomes under deadlines. *Anim. Cogn.* 20, 473–484

55. Tosun, T. *et al.* (2016) Mice plan decision strategies based on previously learned time intervals, locations, and probabilities. *Proc. Natl. Acad. Sci. U. S. A.* 113, 787–792

56. Swanton, D.N. *et al.* (2009) Averaging of temporal memories by rats. *J. Exp. Psychol. Anim. Behav. Process.* 35, 434

57. Li, Y. and Dudman, J.T. (2013) Mice infer probabilistic models for timing. *Proc. Natl. Acad. Sci. U. S. A.* 110, 17154–17159

58. Molet, M. and Miller, R.R. (2014) Timing: an attribute of associative learning. *Behav. Process.* 101, 4–14

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

60. Howard, M.W. *et al.* (2015) A distributed representation of internal time. *Psychol. Rev.* 122, 24–53

61. Molet, M. *et al.* (2012) When does integration of independently acquired temporal relationships take place? *J. Exp. Psychol. Anim. Behav. Process.* 38, 369–380

62. Daw, N.D. *et al.* (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* 8, 1704–1711

63. Kheifets, A. *et al.* (2017) Theoretical implications of quantitative properties of interval timing and probability estimation in mouse and rat. *J. Exp. Anal. Behav.* 108, 39–72

64. Knill, D.C. and Pouget, A. (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719

65. Rao, R.P. (2010) Decision making under uncertainty: a neural model based on partially observable Markov decision processes. *Front. Comput. Neurosci.* 4, 146

66. Lak, A. *et al.* (2017) Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Curr. Biol.* 27, 821–832

67. Fiorillo, C.D. *et al.* (2008) The temporal precision of reward prediction in dopamine neurons. *Nat. Neurosci.* 11, 966–973

68. Gershman, S.J. *et al.* (2015) Discovering latent causes in reinforcement learning. *Curr. Opin. Behav. Sci.* 5, 43–50

69. Gershman, S.J. and Daw, N.D. (2017) Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annu. Rev. Psychol.* 68, 101–128

70. Roach, N.W. *et al.* (2017) Generalization of prior information for rapid Bayesian time estimation. *Proc. Natl. Acad. Sci. U. S. A.* 114, 412–417

71. Acerbi, L. *et al.* (2012) Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS Comput. Biol.* 8, e1002771

72. Jazayeri, M. and Shadlen, M.N. (2010) Temporal context calibrates interval timing. *Nat. Neurosci.* 13, 1020

73. Lejeune, H. and Wearden, J.H. (2009) Vierordt's the experimental study of the time sense (1868) and its legacy. *Eur. J. Cogn. Psychol.* 21, 941–960

74. Shi, Z. *et al.* (2013) Bayesian optimization of time perception. *Trends Cogn. Sci.* 17, 556–564

75. Collins, A.G.E. and Frank, M.J. (2016) Motor demands constrain cognitive rule structures. *PLoS Comput. Biol.* 12, e1004785

76. Harlow, H.F. (1949) The formation of learning sets. *Psychol. Rev.* 56, 51–65

77. Meck, W.H. and Church, R.M. (1982) Abstraction of temporal attributes. *J. Exp. Psychol. Anim. Behav. Process.* 8, 226–243

78. Meegan, D.V. *et al.* (2000) Motor timing learned without motor training. *Nat. Neurosci.* 3, 860–862

79. Nagarajan, S.S. *et al.* (1998) Practice-related improvements in somatosensory interval discrimination are temporally specific but generalize across skin location, hemisphere, and modality. *J. Neurosci.* 18, 1559–1570

80. Wright, B.A. *et al.* (1997) Learning and generalization of auditory temporal–interval discrimination in humans. *J. Neurosci.* 17, 3956–3963

81. Howard, M.W. *et al.* (2015) Efficient neural computation in the Laplace domain. In *Proceedings of the 2015th International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches* (Vol. 1583), pp. 61–68, CEUR-WS.org

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

83. Petter, E.A. and Merchant, H. (2016) Temporal processing by intrinsic neural network dynamics. *Timing Time Percept.* 4, 399–410

84. Murray, J.M. (2017) Learning multiple variable-speed sequences in striatum via cortical tutoring. *eLife* 6, e26084

85. Meck, W.H. and Ivry, R.B. (2016) Time in perception and action. *Curr. Opin. Behav. Sci.* 8, vi–x

86. Spanjaard, O. and Weng, P. (2013) Markov decision processes with functional rewards. In *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, pp. 269–280, Springer

87. Church, R.M. *et al.* (1994) Application of scalar timing theory to individual trials. *J. Exp. Psychol. Anim. Behav. Process.* 20, 135–155

88. Sutton, R.S. and Barto, A.G. (1998) *Reinforcement Learning: An Introduction,* MIT Press

89. Bermudez, M.A. and Schultz, W. (2014) Timing in reward and decision processes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369, 20120468

90. Barter, J.W. *et al.* (2015) Beyond reward prediction errors: the role of dopamine in movement kinematics. *Front. Integr. Neurosci.* 9, 39

91. Kool, W. *et al.* (2018) Planning complexity registers as a cost in metacontrol. *J. Cogn. Neurosci.* 18, 1–14

92. Sutton, R.S. (1991) . In *Reinforcement Learning Architectures for Animats. From Animals to Animats* (. In *Reinforcement Learning Architectures for Animats. From Animals to Animats* (Meyer, J.-A. and Wilson, S.W., eds), pp. 288–296, MIT Press

93. Kaelbling, L.P. *et al.* (1998) Planning and acting in partially observable stochastic domains. *Artif. Intell.* 101, 99–134

94. Nassar, M.R. *et al.* (2010) An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J. Neurosci.* 30, 12366–12378