frontiers in
COMPUTATIONAL NEUROSCIENCE

HYPOTHESIS AND THEORY ARTICLE
published: 09 January 2014
doi: 10.3389/fncom.2013.00194

Time representation in reinforcement learning models of the basal ganglia

Samuel J. Gershman[1]*, Ahmed A. Moustafa[2] and Elliot A. Ludvig[3,4]

[1] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
[2] School of Social Sciences and Psychology, Marcs Institute for Brain and Behaviour, University of Western Sydney, Sydney, NSW, Australia
[3] Princeton Neuroscience Institute and Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA
[4] Department of Psychology, University of Warwick, Coventry, UK

Edited by:
Hagai Bergman, The Hebrew University- Hadassah Medical School, Israel

Reviewed by:
Yoram Burak, Hebrew University, Israel
Daoyun Ji, Baylor College of Medicine, USA

***Correspondence:**
Samuel J. Gershman, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Room 46-4053, 77 Massachusetts Ave., Cambridge, MA 02139, USA
e-mail: sjgershm@mit.edu

Reinforcement learning (RL) models have been influential in understanding many aspects of basal ganglia function, from reward prediction to action selection. Time plays an important role in these models, but there is still no theoretical consensus about what kind of time representation is used by the basal ganglia. We review several theoretical accounts and their supporting evidence. We then discuss the relationship between RL models and the timing mechanisms that have been attributed to the basal ganglia. We hypothesize that a single computational system may underlie both RL and interval timing—the perception of duration in the range of seconds to hours. This hypothesis, which extends earlier models by incorporating a time-sensitive action selection mechanism, may have important implications for understanding disorders like Parkinson's disease in which both decision making and timing are impaired.

Keywords: reinforcement learning, basal ganglia, dopamine, interval timing, Parkinson's disease

INTRODUCTION

Computational models of reinforcement learning (RL) have had a profound influence on the contemporary understanding of the basal ganglia (Joel et al., 2002; Cohen and Frank, 2009). The central claim of these models is that the basal ganglia are organized to support prediction, learning and optimization of long-term reward. While this claim is now widely accepted, RL models have had little to say about the extensive research implicating the basal ganglia in interval timing—the perception of duration in the range of seconds to hours (Buhusi and Meck, 2005; Jones and Jahanshahi, 2009; Merchant et al., 2013). However, this is not to say that time is ignored by these models—on the contrary, time representation has been a pivotal issue in RL theory, particularly with regard to the role of dopamine (Suri and Schultz, 1999; Daw et al., 2006; Ludvig et al., 2008; Nakahara and Kaveri, 2010; Rivest et al., 2010).

In this review, we attempt a provisional synthesis of research on RL and interval timing in the basal ganglia. We begin by briefly reviewing RL models of the basal ganglia, with a focus on how they represent time. We then summarize the key data linking the basal ganglia with interval timing, drawing connections between computational approaches to timing and their relationship to RL models. Our central thesis is that by incorporating a time-sensitive action selection mechanism into RL models, a single computational system can support both RL and interval timing. This unified view leads to a coherent interpretation of decision making and timing deficits in Parkinson's disease.

REINFORCEMENT LEARNING MODELS OF THE BASAL GANGLIA

RL models characterize animals as agents that seek to maximize future reward (for reviews, see Maia, 2009; Niv, 2009; Ludvig et al., 2011). To do so, animals are assumed to generate a prediction of future reward and select actions according to a policy that maximizes that reward. More formally, suppose that at time t an agent occupies a state s_t (e.g., the agent's location or the surrounding stimuli) and receives a reward r_t. The agent's goal is to predict the *expected discounted future return*, or *value*, of visiting a sequence of states starting in state s_t (Sutton and Barto, 1998):

$$V(S_t) = E\left[\sum_{k=0} \gamma^k r_{t+k}\right], \quad (1)$$

where γ is a parameter that discounts distal rewards relative to proximal rewards, and E denotes an average over possibly stochastic sequences of states and rewards.

Typically, a state s_t is described by a set of D features, $\{x_t(1), \ldots, x_t(D)\}$, encoding sensory and cognitive aspects of an animal's current experience. Given this state representation, the value can be approximated by a weighted combination of the features:

$$\hat{V}(s_t) = \sum_d w_t(d)\, x_t(d)$$

where $\hat{V}$ is an estimate of the true value V. According to RL models of the basal ganglia, these features are represented by cortical

inputs to the striatum, with the striatum itself encoding the estimated value (Maia, 2009; Niv, 2009; Ludvig et al., 2011). The strengths of these corticostriatal synapses are represented by a set of weights $\{w_t(1), \ldots, w_t(D)\}$.

These weights can be learned through a simple algorithm known as *temporal-difference (TD) learning*, which adjusts the weights on each time step based on the difference between received and predicted reward:

$$w_{t+1}(d) = w_t(d) + \alpha \delta_t e_t(d),$$

where α is a learning rate and δ_t is a prediction error defined as:

$$\delta_t = r_t + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t).$$

The *eligibility trace* $e_t(d)$ is updated according to:

$$e_{t+1}(d) = \gamma \lambda e_t(d) + x_t(d),$$

where λ is a decay parameter that determines the plasticity window of recent stimuli. The TD algorithm is a computationally efficient method that is known to converge to the true value function [see Equation (1) above] with enough experience and adequate features (Sutton and Barto, 1998).

The importance of this algorithm to neuroscience lies in the fact that the firing of midbrain dopamine neurons conforms remarkably well to the theoretical prediction error (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997; though see Redgrave et al., 2008 for a critique). For example, dopamine neurons increase their firing upon the delivery of an unexpected reward and pause when an expected reward is omitted (Schultz et al., 1997). The role of prediction errors in learning is supported by the observation that plasticity at corticostriatal synapses is gated by dopamine (Reynolds and Wickens, 2002; Steinberg et al., 2013), as well as a large body of behavioral evidence (Rescorla and Wagner, 1972; Sutton and Barto, 1990; Ludvig et al., 2012).

A fundamental question facing RL models is the choice of feature representation. Early applications of TD learning to the dopamine system assumed what is known as the *complete serial compound* (CSC; Moore et al., 1989; Sutton and Barto, 1990; Montague et al., 1996; Schultz et al., 1997), which represents every time step following stimulus onset as a separate feature. Thus, the first feature has a value of 1 for the first time step and 0 for all other time steps, the second feature has a value of 1 for the second time step and 0 for all other time steps, and so on. This CSC representation assumes a perfect clock, whereby the brain always knows exactly how many time steps have elapsed since stimulus onset.

The CSC is effective at capturing several salient aspects of the dopamine response to cued reward. A number of authors (e.g., Daw et al., 2006; Ludvig et al., 2008), however, have pointed out aspects of the dopamine response that appear inconsistent with the CSC. For example, the CSC predicts a large, punctate negative prediction error when an expected reward is omitted; the actual decrease in dopamine response is relatively small and temporally extended (Schultz et al., 1997; Bayer et al., 2007). Another problem with the CSC is that it predicts a large negative prediction error at the usual reward delivery time when a reward is delivered early. Contrary to this prediction, Hollerman and Schultz (1998) found that early reward evoked a large response immediately after the unexpected reward, but showed little change from baseline at the usual reward delivery time.

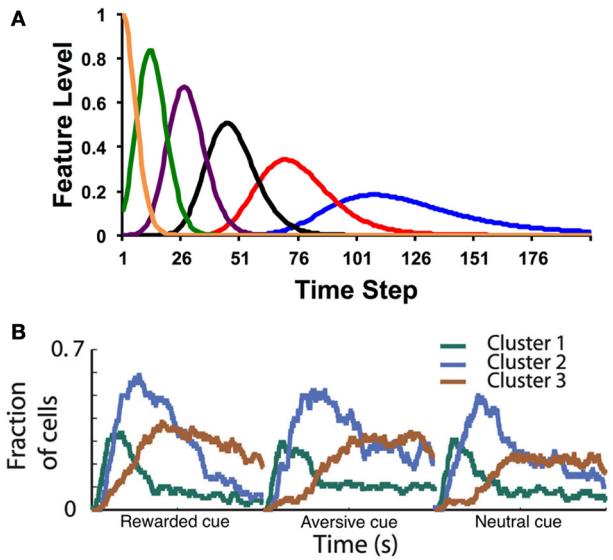

FIGURE 1 | (A) Example microstimuli. Each microstimulus acts as a dynamic feature, going up and then coming back down with a characteristic time course after the stimulus. **(B)** Fraction of medium spiny neurons recorded in the putamen with a significant response (both increases and decreases in firing rate) to cue events. In this task, a visual cue (fractal image) is presented at time 0, followed by an outcome (food, air puff, or neutral sound) 2 s later. Clusters correspond to groups of neurons with similar response profiles (see Adler et al., 2012 for details). Time courses are shown separately for cues that predict rewarding (food), aversive (air puff), or neutral (sound) outcomes. Note the similarity in the population response to the postulated time course of the theoretical microstimuli. Figure adapted from Adler et al. (2012) with permission from the publisher.

It is possible that these mismatches between theory and data reflect problems with a number of different theoretical assumptions. Indeed, several theoretical assumptions have been questioned by recent research (see Niv, 2009). We focus here on alternative time representations as one potential response to the findings mentioned above.

We will discuss two of these alternatives (see also Suri and Schultz, 1999; Nakahara and Kaveri, 2010; Rivest et al., 2010): (1) the microstimulus representation and (2) states with variable durations (a *semi-Markov* formalism) and only partial observability. For the former, Ludvig et al. (2008) proposed that when a stimulus is presented, it leaves a slowly decaying memory trace, which is encoded by a series of temporal receptive fields. Each feature (or "microstimulus") $x_t(d)$ represents the proximity between the trace and the center of the receptive field, producing a spectrum of features that vary with time, as illustrated in **Figure 1A**. Specifically, Ludvig et al. endowed each stimulus with microstimuli of the following form:

$$x_t(d) = \frac{y_t}{\sqrt{\sigma^2}} \exp\left(-\frac{\left(y_t - \frac{d}{D}\right)^2}{2\sigma^2}\right)$$

where D is the number of microstimuli, σ^2 controls the width of each receptive field, and y_t is the stimulus trace strength, which was set to 1 at stimulus onset and decreased exponentially with a decay rate of 0.985 per time step. Both cues and rewards elicit their own set of microstimuli. This feature representation is plugged into the TD learning equations described above.

The microstimulus representation is a temporally smeared version of the CSC: whereas in the CSC each feature encodes a single time point, in the microstimulus representation each feature encodes a temporal range (see also Grossberg and Schmajuk, 1989; Machado, 1997). With the CSC, as time elapses after a stimulus, there is one, unique feature active at each time point. Learned weights for that time point therefore all accrue to that one feature. In contrast, at any time point, a subset of the microstimuli is differentially activated. These serve as the features that can be used to generate a prediction of upcoming reward (values). Note how the temporal precision of the microstimuli decreases with time following stimulus onset, so that later microstimuli are more dispersed than earlier microstimuli.

Recent data from Adler et al. (2012) have provided direct evidence for microstimuli in the basal ganglia (**Figure 1B**). Recording from the putamen while a monkey was engaged in a classical conditioning task, Adler et al. found clusters of medium spiny neurons with distinct post-stimulus time courses (for both cues and outcomes). As postulated by Ludvig et al. (2008), the peak response time varied across clusters, with long latency peaks (i.e., late microstimuli) associated with greater dispersion. Recording from the caudate nucleus, Jin et al. (2009) also found clusters of neurons that encode time-stamps of different events. These neurons carry sufficient information to decode time from the population response. Early time points are decodable with higher fidelity compared to late time points, as would be expected if the dispersion of temporal receptive fields increases with latency.

A different solution to the limitations of the CSC was suggested by Daw et al. (2006). They proposed that dopaminergic prediction errors reflect a state space that is *partially observable* and *semi-Markov*. The partial-observability assumption means that the underlying state is inferred from sensory data (cues and rewards), rather than using the features as a proxy for the state. Thus, prediction errors are computed with respect to a *belief state*, a set of features encoding the probabilistic inference about the hidden state. The semi-Markov assumption means that each state is occupied for a random amount of time before transitioning. In the simulations of Daw et al. (2006), only two states were postulated: an interstimulus interval (ISI) state and an intertrial interval (ITI) state. Rewards are delivered upon transition from the ISI to the ITI state, and cues occur upon transition from the ITI to the ISI state. The learning rule in this model is more complex than the standard TD learning rule [which is used by the Ludvig et al. (2008) model]; however, the core idea of learning from prediction errors is preserved in this model.

It is instructive to compare how these two models account for the data on early reward presented by Hollerman and Schultz (1998). In the Ludvig et al. (2008) model, the weights for all the microstimuli are updated after every time step: the late microstimuli associated with the cue (i.e., those centered around the time of reward delivery) accrue positive weights, even after the time of reward delivery (a consequence of the temporal smearing). These post-reward positive predictions generate a negative prediction error, causing the early microstimuli associated with the reward to accrue negative weights. When reward is presented early, the net prediction is close to zero, because the positive weights on the late cue microstimuli compete with the negative weights on the early reward microstimuli. This interaction produces a negligible negative prediction error, consistent with the data of Hollerman and Schultz (1998). The account of Daw et al. (2006) is conceptually different: when reward is presented early, the model infers that a transition to the ITI state has occurred early, and consequently no reward is expected.

Thus far, we have discussed time representations in the service of RL and their implications for the timing of the dopamine response during conditioning. What do RL models have to say about interval timing *per se*? We will argue below that these are not really separate problems: interval timing tasks can be viewed fundamentally as RL tasks. Concomitantly, the role of dopamine and the basal ganglia in interval timing can be understood in terms of their computational contributions to RL. To elaborate this argument, we need to first review some of the relevant theory and data linking interval timing with the basal ganglia.

TIME REPRESENTATION IN THE BASAL GANGLIA: DATA AND THEORY

The role of the basal ganglia and dopamine in interval timing has been studied most extensively in the context of two procedures: the peak procedure (Catania, 1970; Roberts, 1981) and the bisection procedure (Church and Deluty, 1977). The peak procedure consists of two trial types: on fixed-interval trials, the subject is rewarded if a response is made after a fixed duration following cue presentation. On probe trials, the cue duration is extended, and no reward is delivered for responding. **Figure 2A** shows a typical response curve on probe trials: on average, the response rate peaks around the time of food presentation (20 or 40 s in the figure) is ordinarily available and then decreases. The peak time (a measure of the animal's interval estimate) is the time at which the response rate is maximal.

The other two curves in **Figure 2A** illustrate the standard finding that drugs (or genetic manipulations) that increase dopamine transmission, such as methamphetamine, shift the response curve leftward (Maricq et al., 1981; Matell et al., 2004, 2006; Cheng et al., 2007; Balci et al., 2010), whereas drugs that decrease dopamine transmission shift the response curve rightward (Drew et al., 2003; Macdonald and Meck, 2005).

In the bisection procedure, subjects are trained to respond differentially to short and long duration cues. Unreinforced probe trials with cue durations between these two extremes are occasionally presented. On these trials, typically, a psychometric curve is produced with greater selection of the long option (i.e.,

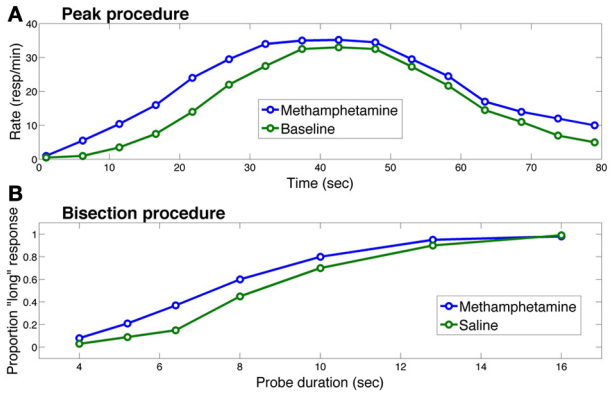

FIGURE 2 | Effects of methamphetamine on timing in (A) the peak procedure and (B) the bisection procedure in rats. Each curve in **(A)** represents response rate as a function of time, where time 0 corresponds to the trial onset. The methamphetamine curve corresponds to sessions in which rats were injected with methamphetamine; the baseline curve corresponds to sessions in which rats did not receive an injection. Each curve in **(B)** represents the proportion of trials on which the rat chose the "long" option as a function of probe cue duration. The saline curve corresponds to sessions in which the rat received a saline injection. In both procedures, methamphetamine leads to overestimation of the elapsing interval, producing early responding in the peak procedure and more "long" responses in the bisection procedure. Figure replotted from Maricq et al. (1981).

the option reinforced following long duration cues) with longer probes and greater selection of the short option with shorter probes and a gradual shift between the two (see **Figure 2B**). The indifference point or point of subjective equality is typically close to the geometric mean of the two anchor durations (Church and Deluty, 1977). Similar to the peak procedure, in the bisection procedure, **Figure 2B** shows how dopamine agonists usually produce a leftward shift in the psychometric curve—i.e., more "short" responses, whereas dopamine antagonists produce the opposite pattern (Maricq et al., 1981; Maricq and Church, 1983; Meck, 1986; Cheng et al., 2007). Under some circumstances, however, dopamine agonists induce temporal dysregulation with an overall flattening of the response curve and no shift in preference or peak times (e.g., Odum et al., 2002; McClure et al., 2005; Balci et al., 2008).

The most influential interpretation of these findings draws upon the class of pacemaker-accumulator models (Gibbon et al., 1997), according to which a pacemaker (an "internal clock") emits pulses that are accumulated by a counter to form a representation of subjective time intervals. The neurobiological implementation of this scheme might rely on populations of oscillating neurons (Miall, 1989; Matell and Meck, 2004), integration of ramping neural activity (Leon and Shadlen, 2003; Simen et al., 2011), or intrinsic dynamics of a recurrent network (Buonomano and Laje, 2010). Independent of the neural implementation, the idea is that drugs that increase dopamine speed up the internal clock, while drugs that decrease dopamine slow the internal clock down.

This interpretation is generally consistent with the findings from studies of patients with Parkinson's disease (PD), who have chronically low striatal dopamine levels. When off medication, these patients tend to underestimate the length of temporal intervals in verbal estimation tasks; dopaminergic medication alleviates this underestimation (Pastor et al., 1992; Lange et al., 1995). It should be noted, however, that some studies have found normal time perception in PD (Malapani et al., 1998; Spencer and Ivry, 2005; Wearden et al., 2008), possibly due to variations in disease severity (Artieda et al., 1992).

Pacemaker-accumulator models have been criticized on a number of grounds, such as lack of parsimony, implausible neurophysiological assumptions, and incorrect behavioral predictions (Staddon and Higa, 1999, 2006; Matell and Meck, 2004; Simen et al., 2013). Moreover, while the pharmacological data are generally consistent with the idea that dopamine modulates the speed of the internal clock, these data may also be consistent with other interpretations. One important alternative is the class of "distributed elements" models, which postulate a representation of time that is distributed over a set of elements; these elements come in various flavors, such as "behavioral states" (Machado, 1997), a cascade of leaky integrators (Staddon and Higa, 1999, 2006; Shankar and Howard, 2012), or spectral traces (Grossberg and Schmajuk, 1989). The effects of dopaminergic drugs might be explicable in terms of systematic changes in the pattern of activity across the distributed elements (see, for example, Grossberg and Schmajuk, 1989).

In fact, the microstimulus model of Ludvig et al. (2008) can be viewed as a distributed elements model embedded within the machinery of RL. This connection suggests a more ambitious theoretical synthesis: can we understand the behavioral and neurophysiological characteristics of interval timing in terms of RL?

TOWARD A UNIFIED MODEL OF REINFORCEMENT LEARNING AND TIMING

One suggestive piece of evidence for how RL models and interval timing can be integrated comes from the study of Fiorillo et al. (2008); (see also Kobayashi and Schultz, 2008). They trained monkeys on a variation of the peak procedure with classical contingencies (i.e., water was delivered independent of responding) while recording from dopamine neurons in the substantia nigra and ventral tegmental area with five different intervals spanning from 1 to 16 s. As shown in **Figure 3A**, they found that the dopamine response to the reward increased with the interval, and the dopamine response to the cue decreased with the interval.

Whereas the response to the cue can be explained in terms of temporal discounting, the response to the reward should not (according to the CSC representation) depend on the cue-reward interval. The perfect timing inherent in the CSC representation means that the reward can be equally well predicted at all time points. Thus, there should be no reward-prediction error, and no phasic dopamine response, at the time of reward regardless of the cue-reward interval. Alternatively, the dopamine response to reward can be understood as reflecting increasing uncertainty in the temporal prediction. **Figure 3B** shows how, using the microstimulus TD model as defined as in Ludvig et al. (2008), there is indeed an increase in the simulated reward prediction error as a function of interval. In the model, with longer intervals,

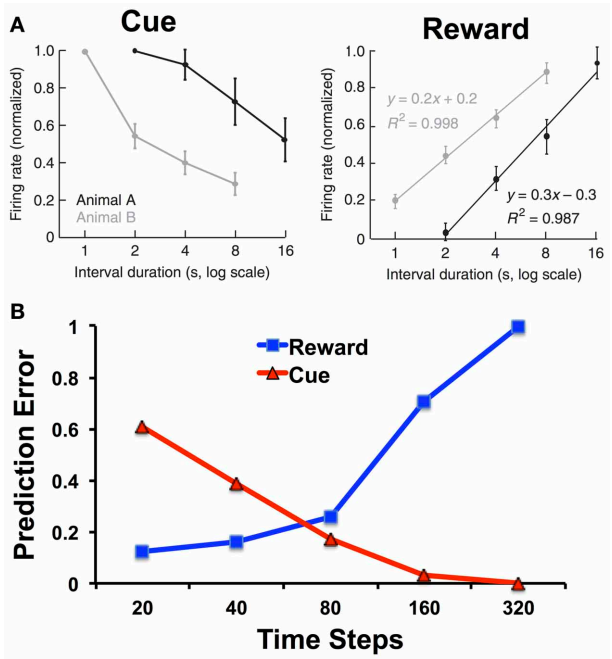

FIGURE 3 | (A) Firing rates of dopamine neurons in monkeys to cues and rewards as a function of cue-reward interval duration. Adapted from Fiorillo et al. (2008) with permission from the publisher. **(B)** Prediction errors to the reward and cue as a function of the interval duration for the microstimulus model. The simulations of the microstimulus model were run for 100 trials and used the same parameters specified in Ludvig et al. (2008): $\lambda = 0.95$, $\alpha = 0.01$, $\gamma = 0.98$, $D = 50$, $\sigma = 0.08$. We treated 20 time steps as a unit of 1 s, and each trial was separated by an interval of 500 time steps. Note the logarithmic scale on the x-axis.

the reward predictions are less temporally precise, and greater prediction errors persist upon reward receipt.

Interval timing procedures, such as the peak procedure, add an additional nuance to this problem by introducing instrumental contingencies. Animals must now not only predict the timing of reward, but also learn when to respond. To analyze this problem in terms of RL, we need to augment the framework introduced earlier to have actions. There are various ways to accomplish this (see Sutton and Barto, 1998). The Actor-Critic architecture (Houk et al., 1995; Joel et al., 2002) is one of the earliest and most influential approaches; it postulates a separate "actor" mechanism that probabilistically chooses an action a_t given the current state s_t. The action probabilities $p(s_t|a_t) \propto \exp\{f(s_t|a_t)\}$ are updated according to:

$$f(s_t|a_t) \leftarrow f(s_t|a_t) + \eta\delta_t[1 - p(s_t|a_t)],$$

where η is a learning rate parameter and δ_t is the is the prediction error defined earlier. The value estimation system plays the role of a "critic" that teaches the actor how to modify its action selection probabilities so as to reduce prediction errors.

When combined with the microstimulus representation, the actor-critic architecture naturally gives rise to timing behavior: in the peak procedure, on average, responding will tend to increase toward the expected reward time and decrease thereafter (see **Figure 2**). Importantly, the late microstimuli are less temporally precise than the early microstimuli, in the sense that their responses are more dispersed over time. As a consequence, credit for late rewards is assigned to a larger number of microstimuli. Under the assumption that response rate is proportion to predicted value, this dispersion of credit causes the timing of actions to be more spread out around the time of reward as the length of the interval increases, one of the central empirical regularities in timing behavior (Gibbon, 1977; see also Ludvig et al., 2012 for an exploration of this property in classical conditioning). As described above, an analog of this property has also been observed in the firing of midbrain dopamine neurons: response to reward increases linearly with the logarithm of the stimulus-reward interval, consistent with the idea that prediction errors are being computed with respect to a value signal whose temporal precision decreases over time (Fiorillo et al., 2008). To the best of our knowledge, pacemaker-accumulator models cannot account for the results presented in **Figure 3**, because they do not have reward-prediction errors in their machinery. Instead, they collect a distribution of past cue-reward intervals and draw from that distribution to create an estimate of the time to reward (e.g., Gibbon et al., 1984).

The partially observable semi-Markov model of Daw et al. (2006) can account for the findings of Fiorillo et al. (2008), but this account deviates from the normative RL framework. Daw et al. use an external timing signal with "scalar noise" (cf. Gibbon et al., 1984), implemented by adding Gaussian noise to the timing signal with standard deviation proportional to the interval. Scalar noise induces larger-magnitude prediction errors with increasing delays. However, these prediction errors are symmetric around 0 and hence cancel out on average. To account for the effects of cue-reward interval on the dopamine response, Daw et al. assume that negative prediction errors are rectified (see Bayer and Glimcher, 2005), resulting in a positive skew of the prediction error distribution. **Figure 4** shows how this asymmetric rectification results in average prediction errors that are increasingly positive for longer intervals. Note that rectification is not an intrinsic part of the RL framework, and in fact compromises the convergence of TD to the true value function. One potential solution to this problem is to posit a separate physiological channel for the signaling of negative prediction errors, possibly via serotonergic activity (Daw et al., 2002).

The microstimulus actor-critic model can also explain the effects of dopamine manipulations and Parkinson's disease. The key additional assumption is that early microstimuli (but not later ones) are primarily represented by the striatum. Timing in the milliseconds to seconds range depends on D2 receptors in the dorsal striatum (Rammsayer, 1993; Coull et al., 2011), suggesting that this region represents early microstimuli (whereas late microstimuli may be represented by other neural substrates, such as the hippocampus; see Ludvig et al., 2009). Because the post-synaptic effect of dopamine at D2 receptors is inhibitory, D2 receptor antagonists increase the firing of striatal neurons expressing D2 receptors, which mainly occur in the indirect or "NoGo" pathway and exert a suppressive influence on striatal output (Gerfen, 1992). Thus, the ultimate effect of D2 receptor

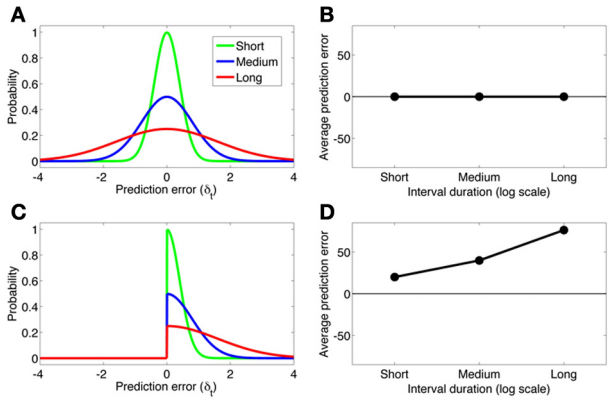

FIGURE 4 | (A) The distribution of prediction errors in the semi-Markov TD model (Daw et al., 2006) for three logarithmically spaced cue-reward interval durations (short, medium, long). The timing signal is corrupted by zero-mean Gaussian noise with standard deviation proportional to the interval duration. **(B)** The average prediction error under each distribution. A symmetric prediction error distribution implies that the average prediction error is invariant to the interval duration. **(C,D)** Same format as panels A and B, but where the prediction error distribution is rectified slightly below 0. This asymmetric distribution produces average prediction errors that scale approximately linearly with the log interval duration.

antagonists is to reduce striatal output, thereby attenuating the influence of early microstimuli on behavior. As a result, predictions of the upcoming reward will be biased later, and responses will occur later than usual (e.g., in the peak procedure). This fits with the observation that the rightward shift (overestimation) of estimated time following dopamine antagonist administration is proportional to the drug's binding affinity for D2 receptors (Meck, 1986). In contrast, dopamine agonists lead to a selective enhancement of the early microstimuli, producing earlier than usual responding (see **Figure 2A**).

A similar line of reasoning can explain some of the timing deficits in Parkinson's disease. The nigrostriatal pathway (the main source of dopamine to the dorsal striatum) is compromised in Parkinson's disease, resulting in reduced striatal dopamine levels. Because D2 receptors have a higher affinity for dopamine, Parkinson's disease leads to the predominance of D2-mediated activity and hence reduced striatal output (Wiecki and Frank, 2010). Our model thus predicts a rightward shift of estimated time, as is often observed experimentally (see above).

The linking of early microstimuli with the striatum in the model also leads to the prediction that low striatal dopamine levels will result in poorer learning of fast responses (which depend on the early microstimuli). In addition, responding will in general be slowed because the learned weights to the early microstimuli will be weak relative to those of late microstimuli. As a result, our model clearly predicts poorer learning of fast responses in Parkinson's disease. A study of temporal decision making in Parkinson's patients fits with this prediction (Moustafa et al., 2008). Patients were trained to respond at different latencies to a set of cues, with slow responses yielding more reward in an "increasing expected value" (IEV) condition and fast responses yielding more reward in a "decreasing expected value" (DEV) condition. It was found that the performance of medicated patients was better in the DEV condition, while performance of non-medicated patients was better in the IEV condition. If non-medicated patients have a paucity of short-timescale microstimuli (due to low striatal dopamine levels), then the model correctly anticipates that these patients will be impaired at learning about early events relative to later events.

Recently, Foerde et al. (2013) found that Parkinson's patients are impaired in learning from immediate, but not delayed, feedback in a probabilistic decision making task. This finding is also consistent with the idea that these patients lack a representational substrate for early post-stimulus events. Interestingly, they also found that patients with lesions of the medial temporal lobe show the opposite pattern, possibly indicating that this region (and in particular the hippocampus) is important for representing late microstimuli, as was suggested earlier by Ludvig et al. (2009). An alternative possibility is that the striatum and hippocampus both represent early and late microstimuli, but the stimulus trace decays more quickly in the striatum than in the hippocampus (Bornstein and Daw, 2012), which would produce a more graded segregation of temporal sensitivity between the two regions.

CONCLUSION

Timing and RL have for the most part been studied separately, giving rise to largely non-overlapping computational models. We have argued here, however, that these models do in fact share some important commonalities and reconciling them may provide a unified explanation of many behavioral and neural phenomena. While in this brief review we have only sketched such a synthesis, our goal is to plant the seeds for future theoretical unification.

One open question concerns how to reconcile the disparate theoretical ideas about time representation that were described in this paper. Our synthesis proposed a central role for a distributed elements representation of time such as the microstimuli of Ludvig et al. (2008). Could a representation deriving from the semi-Markov or pacemaker-accumulator models be used instead? This may be possible, but there are several reasons to prefer the microstimulus representation. First, microstimuli lend themselves naturally to the linear function approximation architecture that has been widely used in RL models of the basal ganglia. In contrast, the semi-Markov model requires additional computational machinery, and it is not obvious how to incorporate the pacemaker-accumulator model into RL theory. Second, the semi-Markov model accounts for the relationship between temporal precision and interval length at the expense of deviating from the normative RL framework. Third, as we noted earlier, pacemaker-accumulator models have a number of other weaknesses (see Staddon and Higa, 1999, 2006; Matell and Meck, 2004; Simen et al., 2013), such as lack of parsimony, implausible neurophysiological assumptions, and incorrect behavioral predictions. Nonetheless, it will be interesting to explore what aspects of these models can be successfully incorporated into the next generation of RL models.

ACKNOWLEDGMENTS

We thank Marc Howard and Nathaniel Daw for helpful discussions. Samuel J. Gershman was supported by IARPA via DOI

contract D10PC2002 and by a postdoctoral fellowship from the MIT Intelligence Initiative. Ahmed A. Moustafa is partially supported by a 2013 internal UWS Research Grant Scheme award P00021210. Elliot A. Ludvig was partially supported by NIH Grant #P30 AG024361 and the Princeton Pyne Fund.

REFERENCES

Adler, A., Katabi, S., Finkes, I., Israel, Z., Prut, Y., and Bergman, H. (2012). Temporal convergence of dynamic cell assemblies in the striato-pallidal network. *J. Neurosci.* 32, 2473–2484. doi: 10.1523/JNEUROSCI.4830-11.2012

Artieda, J., Pastor, M. A., Lacruz, F., and Obeso, J. A. (1992). Temporal discrimination is abnormal in Parkinson's disease. *Brain* 115, 199–210. doi: 10.1093/brain/115.1.199

Balci, F., Ludvig, E. A., Abner, R., Zhuang, X., Poon, P., and Brunner, D. (2010). Motivational effects on interval timing in dopamine transporter (DAT) knockdown mice. *Brain Res.* 1325, 89–99. doi: 10.1016/j.brainres.2010.02.034

Balci, F., Ludvig, E. A., Gibson, J. M., Allen, B. D., Frank, K. M., Kapustinski, B. J., et al. (2008). Pharmacological manipulations of interval timing using the peak procedure in male C3H mice. *Psychopharmacology (Berl.)* 201, 67–80. doi: 10.1007/s00213-008-1248-y

Bayer, H. M., and Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47, 129–141. doi: 10.1016/j.neuron.2005.05.020

Bayer, H. M., Lau, B., and Glimcher, P. W. (2007). Statistics of midbrain dopamine neuron spike trains in the awake primate. *J. Neurophysiol.* 98, 1428–1439. doi: 10.1152/jn.01140.2006

Bornstein, A. M., and Daw, N. D. (2012). Dissociating hippocampal and striatal contributions to sequential prediction learning. *Eur. J. Neurosci.* 35, 1011–1023. doi: 10.1111/j.1460-9568.2011.07920.x

Buhusi, C. V., and Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nat. Rev. Neurosci.* 6, 755–765. doi: 10.1038/nrn1764

Buonomano, D. V., and Laje, R. (2010). Population clocks: motor timing with neural dyanmics. *Trends Cogn. Sci.* 14, 520–527. doi: 10.1016/j.tics.2010.09.002

Catania, R. A. (1970). "Reinforcement schedules and psychophysical judgements," in *The Theory of Reinforcement Schedules*, ed W. N. Schoenfeld (New York, NY: Appleton-Century Crofts), 1–42.

Cheng, R. K., Ali, Y. M., and Meck, W. H. (2007). Ketamine "unlocks" the reduced clock-speed effects of cocaine following extended training: evidence for dopamine-glutamate interactions in timing and time perception. *Neurobiol. Learn. Mem.* 88, 149–159. doi: 10.1016/j.nlm.2007.04.005

Church, R. M., and Deluty, H. Z. (1977). The bisection of temporal intervals. *J. Exp. Psychol. Anim. Behav. Process.* 3, 216–228. doi: 10.1037/0097-7403.3.3.216

Cohen, M. X., and Frank, M. J. (2009). Neurocomputational models of basal ganglia function in learning, memory and choice. *Behav. Brain Res.* 199, 141–156. doi: 10.1016/j.bbr.2008.09.029

Coull, J. T., Cheng, R. K., and Meck, W. H. (2011). Neuroanatomical and neurochemical substrates of timing. *Neuropsychopharmacology* 36, 3–25. doi: 10.1038/npp.2010.113

Daw, N. D., Courville, A. C., and Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Comput.* 18, 1637–1677. doi: 10.1162/neco.2006.18.7.1637

Daw, N. D., Kakade, S., and Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Netw.* 15, 603–616. doi: 10.1016/S0893-6080(02)00052-7

Drew, M. R., Fairhurst, S., Malapani, C., Horvitz, J. C., and Balsam, P. D. (2003). Effects of dopamine antagonists on the timing of two intervals. *Pharmacol. Biochem. Behav.* 75, 9–15. doi: 10.1016/S0091-3057(03)00036-4

Fiorillo, C. D., Newsome, W. T., and Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. *Nat. Neurosci.* 11, 966–973. doi: 10.1038/nn.2159

Foerde, K., Race, E., Verfaellie, M., and Shohamy, D. (2013). A role for the medial temporal lobe in feedback-driven learning: evidence from amnesia. *J. Neurosci.* 33, 5698–5704. doi: 10.1523/JNEUROSCI.5217-12.2013

Gerfen, C. R. (1992). The neostriatal mosaic: multiple levels of compartmental organization in the basal ganglia. *Annu. Rev. Neurosci.* 15, 285–320. doi: 10.1146/annurev.ne.15.030192.001441

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychol. Rev.* 84, 279–325. doi: 10.1037/0033-295X.84.3.279

Gibbon, J., Church, R. M., and Meck, W. H. (1984). "Scalar timing in memory," in *Annals of the New York Academy of Sciences: Timing and Time Perception*, Vol. 423, eds J. Gibbon and L. G. Allan (New York, NY: New York Academy of Sciences), 52–77.

Gibbon, J., Malapani, C., Dale, C. L., and Gallistel, C. R. (1997). Towards a neurobiology of temporal cognition: advances and challenges. *Curr. Opin. Neurobiol.* 7, 170–184. doi: 10.1016/S0959-4388(97)80005-0

Grossberg, S., and Schmajuk, N. A. (1989). Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Netw.* 2, 79–102. doi: 10.1016/0893-6080(89)90026-9

Hollerman, J. R., and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* 1, 304–309. doi: 10.1038/1124

Houk, J. C., Adams, J. L., and Barto, A. G. (1995). "A model of how the basal ganglia generate and use neural signals that predict reinforcement," in *Models of information Processing in the Basal Ganglia*, eds J. C. Houk, J. L. Davis, and D. G. Beiser (Cambridge, MA: MIT Press), 249–270.

Jin, D. Z., Fujii, N., and Graybiel, A. M. (2009). Neural representation of time in cortico-basal ganglia circuits. *Proc. Natl. Acad. Sci. U.S.A.* 106, 19156–19161. doi: 10.1073/pnas.0909881106

Joel, D., Niv, Y., and Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw.* 15, 535–547. doi: 10.1016/S0893-6080(02)00047-3

Jones, C. R., and Jahanshahi, M. (2009). The substantia nigra, the basal ganglia, dopamine and temporal processing. *J. Neural Transm. Suppl.* 73, 161–171. doi: 10.1007/978-3-211-92660-4_13

Kobayashi, S., and Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. *J. Neurosci.* 28, 7837–7846. doi: 10.1523/JNEUROSCI.1600-08.2008

Lange, K. W., Tucha, O., Steup, A., Gsell, W., and Naumann, M. (1995). Subjective time estimation in Parkinson's disease. *J. Neural Transm. Suppl.* 46, 433–438.

Leon, M. I., and Shadlen, M. N. (2003). Representation of time by neurons in the posterior parietal cortex of the macaque. *Neuron* 38, 317–327. doi: 10.1016/S0896-6273(03)00185-5

Ludvig, E. A., Bellemare, M. G., and Pearson, K. G. (2011). "A primer on reinforcement learning in the brain: psychological, computational, and neural perspectives," in *Computational Neuroscience for Advancing Artificial Intelligence: Models, Methods and Applications*, eds E. Alonso and E. Mondragon (Hershey, PA: IGI Global), 111–144.

Ludvig, E. A., Sutton, R. S., and Kehoe, E. J. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Comput.* 20, 3034–3054. doi: 10.1162/neco.2008.11-07-654

Ludvig, E. A., Sutton, R. S., and Kehoe, E. J. (2012). Evaluating the TD model of classical conditioning. *Learn. Behav.* 40, 305–319. doi: 10.3758/s13420-012-0082-6

Ludvig, E. A., Sutton, R. S., Verbeek, E. L., and Kehoe, E. J. (2009). A computational model of hippocampal function in trace conditioning. *Adv. Neural Inf. Process. Syst.* 21, 993–1000.

Macdonald, C. J., and Meck, W. H. (2005). Differential effects of clozapine and haloperidol on interval timing in the supraseconds range. *Psychopharmacology (Berl.)* 182, 232–244. doi: 10.1007/s00213-005-0074-8

Machado, A. (1997). Learning the temporal dynamics of behavior. *Psychol. Rev.* 104, 241–265. doi: 10.1037/0033-295X.104.2.241

Maia, T. (2009). Reinforcement learning, conditioning, and the brain: successes and challenges. *Cogn. Affect. Behav. Neurosci.* 9, 343–364. doi: 10.3758/CABN.9.4.343

Malapani, C., Rakitin, B., Levy, R., Meck, W. H., Deweer, B., Dubois, B., et al. (1998). Coupled temporal memories in Parkinson's disease: a dopamine-related dysfunction. *J. Cogn. Neurosci.* 10, 316–331. doi: 10.1162/089892998562762

Maricq, A. V., and Church, R. M. (1983). The differential effects of haloperidol and methamphetamine on time estimation in the rat. *Psychopharmacology (Berl.)* 79, 10–15. doi: 10.1007/BF00433008

Maricq, A. V., Roberts, S., and Church, R. M. (1981). Methamphetamine and time estimation. *J. Exp. Psychol. Anim. Behav. Process.* 7, 18–30. doi: 10.1037//0097-7403.7.1.18

Matell, M. S., Bateson, M., and Meck, W. H. (2006). Single-trials analyses demonstrate that increases in clock speed contribute to the

methamphetamine-induced horizontal shifts in peak-interval timing functions. *Psychopharmacology (Berl.)* 188, 201–212. doi: 10.1007/s00213-006-0489-x

Matell, M. S., King, G. R., and Meck, W. H. (2004). Differential modulation of clock speed by the administration of intermittent versus continuous cocaine. *Behav. Neurosci.* 118, 150–156. doi: 10.1037/0735-7044.118.1.150

Matell, M. S., and Meck, W. H. (2004). Cortico-striatal circuits and interval timing: coincidence detection of oscillatory processes. *Cogn. Brain Res.* 21, 139–170. doi: 10.1016/j.cogbrainres.2004.06.012

McClure, E. A., Saulsgiver, K. A., and Wynne, C. D. L. (2005). Effects of d-amphetamine on temporal discrimination in pigeons. *Behav. Pharmacol.* 16, 193–208. doi: 10.1097/01.fbp.0000171773.69292.bd

Meck, W. H. (1986). Affinity for the dopamine D2 receptor predicts neuroleptic potency in decreasing the speed of an internal clock. *Pharmacol. Biochem. Behav.* 25, 1185–1189. doi: 10.1016/0091-3057(86)90109-7

Merchant, H., Harrington, D. L., and Meck, W. H. (2013). Neural basis of the perception and estimation of time. *Annu. Rev. Neurosci.* 36, 313–336. doi: 10.1146/annurev-neuro-062012-170349

Miall, C. (1989). The storage of time intervals using oscillating neurons. *Neural Comput.* 1, 359–371. doi: 10.1162/neco.1989.1.3.359

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* 16, 1936–1947.

Moore, J. W., Desmond, J. E., and Berthier, N. E. (1989). Adaptively timed conditioned responses and the cerebellum: a neural network approach. *Biol. Cybern.* 62, 17–28. doi: 10.1007/BF00217657

Moustafa, A. A., Cohen, M. X., Sherman, S. J., and Frank, M. J. (2008). A role for dopamine in temporal decision making and reward maximization in Parkinsonism. *J. Neurosci.* 28, 12294–12304. doi: 10.1523/JNEUROSCI.3116-08.2008

Nakahara, H., and Kaveri, S. (2010). Internal-time temporal difference model for neural value-based decision making. *Neural Comput.* 22, 3062–3106. doi: 10.1162/NECO_a_00049

Niv, Y. (2009). Reinforcement learning in the brain. *J. Math. Psychol.* 53, 139–154. doi: 10.1016/j.jmp.2008.12.005

Odum, A. L., Lieving, L. M., and Schaal, D. W. (2002). Effects of D-amphetamine in a temporal discrimination procedure: selective changes in timing or rate dependency? *J. Exp. Anal. Behav.* 78, 195–214. doi: 10.1901/jeab.2002.78-195

Pastor, M. A., Artieda, J., Jahanshahi, M., and Obeso, J. A. (1992). Time estimation and reproduction is abnormal in Parkinson's disease. *Brain* 115, 211–225. doi: 10.1093/brain/115.1.211

Rammsayer, T. H. (1993). On dopaminergic modulation of temporal information processing. *Biol. Psychol.* 36, 209–222. doi: 10.1016/0301-0511(93)90018-4

Redgrave, P., Gurney, K., and Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain Res. Rev.* 58, 322–339. doi: 10.1016/j.brainresrev.2007.10.007

Rescorla, R. A., and Wagner, A. R. (1972). "A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement," in *Classical Conditioning II: Current Research and Theory*, eds A. H. Black and W. F. Prokasy (New York, NY: Appleton-Century Crofts), 64–69.

Reynolds, J. N. J., and Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Netw.* 15, 507–521. doi: 10.1016/S0893-6080(02)00045-X

Rivest, F., Kalaska, J. F., and Bengio, Y. (2010). Alternative time representation in dopamine models. *J. Comput. Neurosci.* 28, 107–130. doi: 10.1007/s10827-009-0191-1

Roberts, S. (1981). Isolation of an internal clock. *J. Exp. Psychol. Anim. Behav. Process.* 7, 242–268. doi: 10.1037/0097-7403.7.3.242

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

Shankar, K. H., and Howard, M. W. (2012). A scale-invariant internal representation of time. *Neural Comput.* 24, 134–193. doi: 10.1162/NECO_a_00212

Simen, P., Balci, F., deSouza, L., Cohen, J. D., and Holmes, P. (2011). A model of interval timing by neural integration. *J. Neurosci.* 31, 9238–9253. doi: 10.1523/JNEUROSCI.3121-10.2011

Simen, P., Rivest, F., Ludvig, E. A., Balci, F., and Killeen, P. (2013). Timescale invariance in the pacemaker-accumulator family of timing models. *Timing Time Percept.* 1, 159–188. doi: 10.1163/22134468-00002018

Spencer, R. M., and Ivry, R. B. (2005). Comparison of patients with Parkinson's disease or cerebellar lesions in the production of periodic movements involving event-based or emergent timing. *Brain Cogn.* 58, 84–93. doi: 10.1016/j.bandc.2004.09.010

Staddon, J. E. R., and Higa, J. J. (1999). Time and memory: towards a pacemaker-free theory of interval timing. *J. Exp. Anal. Behav.* 71, 215–251. doi: 10.1901/jeab.1999.71-215

Staddon, J. E. R., and Higa, J. J. (2006). Interval timing. *Nat. Rev. Neurosci.* 7. doi: 10.1038/nrn1764-c1

Steinberg, E. E., Keiflin, R., Bolvin, J. R., Witten, I. B., Deisseroth, K., and Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* 16, 966–973. doi: 10.1038/nn.3413

Suri, R. E., and Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience* 91, 871–890. doi: 10.1016/S0306-4522(98)00697-6

Sutton, R. S. and Barto, A. G. (1990). "Time-derivative models of Pavlovian reinforcement," in *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, eds M. Gabriel and J. Moore (Cambridge, MA: MIT Press), 497–537.

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

Wearden, J. H., Smith-Spark, J. H., Cousins, R., Edelstyn, N. M., Cody, F. W., and O'Boyle, D. J. (2008). Stimulus timing by people with Parkinson's disease. *Brain Cogn.* 67, 264–279. doi: 10.1016/j.bandc.2008.01.010

Wiecki, T. V., and Frank, M. J. (2010). Neurocomputational models of motor and cognitive deficits in Parkinson's disease. *Prog. Brain Res.* 183, 275–297. doi: 10.1016/S0079-6123(10)83014-6

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 October 2013; accepted: 23 December 2013; published online: 09 January 2014.

Citation: Gershman SJ, Moustafa AA and Ludvig EA (2014) Time representation in reinforcement learning models of the basal ganglia. Front. Comput. Neurosci. 7:194. doi: 10.3389/fncom.2013.00194

This article was submitted to the journal Frontiers in Computational Neuroscience.