TITLE

Application of a unifying reward-prediction error (RPE)-based framework to explain underlying dynamic dopaminergic activity in timing tasks

AUTHORS

Allison E. Hamilos[1] & John A. Assad[1,2]

AFFILIATIONS

[1]Department of Neurobiology, Harvard Medical School, Boston, Massachusetts, 02115, USA.

[2]Istituto Italiano di Tecnologia, Genova, Italy.

Correspondence: A.H. (ahamilos@mit.edu) or J.A.A. (jassad@hms.harvard.edu).

SUMMARY

This manuscript is intended as a theoretical companion to Hamilos *et al.*, 2020[1], in which we examined the role of dopaminergic neurons (DANs) in self-timed movements. In that study, we recorded DAN signals in mice trained to initiate a licking movement after a self-timed delay following a start-timing cue. DAN signals both before the start-timing cue and during the timing interval predicted the timing of movement onset, up to seconds before the movement itself. In particular, dopaminergic signals "ramped up" from the time of the cue to the time of the movement. On a given trial, the slope of the ramping was predictive of when the movement would occur, with steep slope associated with early movement and shallow slope with late movement, reminiscent of a ramp-to-threshold process.

Ramping dopaminergic signals were recently proposed in a theoretical framework that examined temporal-difference learning under resolved state uncertainty (Mikhael *et al.*, 2019[2]; Mikhael & Gershman, 2019[3]; Gershman, 2014[4]). Here, we show that an adapted version of Mikhael *et al.*'s model recapitulates the ramping dopaminergic signaling observed in our self-timed movement task. We also applied the model to results reported in a recent temporal bisection study, in which mice categorized time intervals as relatively short or long compared to a criterion interval (Soares *et al.*, 2016[5]). The model successfully predicted the relative amplitude of dynamic DAN signals observed in the bisection task. These combined results suggest a common neural mechanism that broadly underlies timing behavior: trial-by-trial variation in the rate of the internal "pacemaker," manifested in DAN signals that reflect stretching or compression of the derivative of the subjective value function relative to veridical time. In this view, faster pacemaking is associated with relatively high amplitude dopaminergic signaling, whereas slower pacemaking is associated with relatively low levels of dopaminergic signaling.

MAIN TEXT

Nigrostriatal dopaminergic signaling controls the moment-to-moment decision of when to move

Clues from human movement disorders and pharmacological studies have long suggested a connection between the neurotransmitter dopamine and the timing of movement initiation[3,5-13]. We recently showed that dopaminergic signaling controls the moment-to-moment timing of movements in mice[1]. We recorded dopaminergic signals with fiber photometry in mice executing

a self-timed movement task, in which animals received juice rewards for withholding movement for a proscribed interval (3.3 s) after a start-timing cue and then initiating movement (a first-lick) within a rewarded time window (3.3-7 s, Figure 1a). We observed two aspects of dopaminergic signaling that predicted movement timing: 1) pre-trial baseline signaling of nigrostriatal dopamine neurons (DANs), and, 2) slow "ramping" signals that built up over the course of seconds between the start-timing cue and the self-timed movement. Although self-timed movements occurred with variable timing relative to the start-timing cue[1], DAN signaling rose to about the same level at the moment of movement onset, reminiscent of a ramp-to-threshold process (Figure 1b). DAN signals were not explained by ongoing nuisance movements nor optical artifacts and were best modeled with timing-dependent predictors, including a baseline offset term whose amplitude was proportional to the mouse's timing on the upcoming trial, as well as a stretch feature that encoded percentages of elapsed time between the cue and self-timed movement[1]. DAN ramping activity predicted first-lick time on single trials, independently of trial history, and optogenetic manipulation of DANs bidirectionally shifted movement timing, with activation early-shifting movements versus inhibition late-shifting movements. Together, these results indicate that dopaminergic signaling during self-timing controls the moment of movement onset.

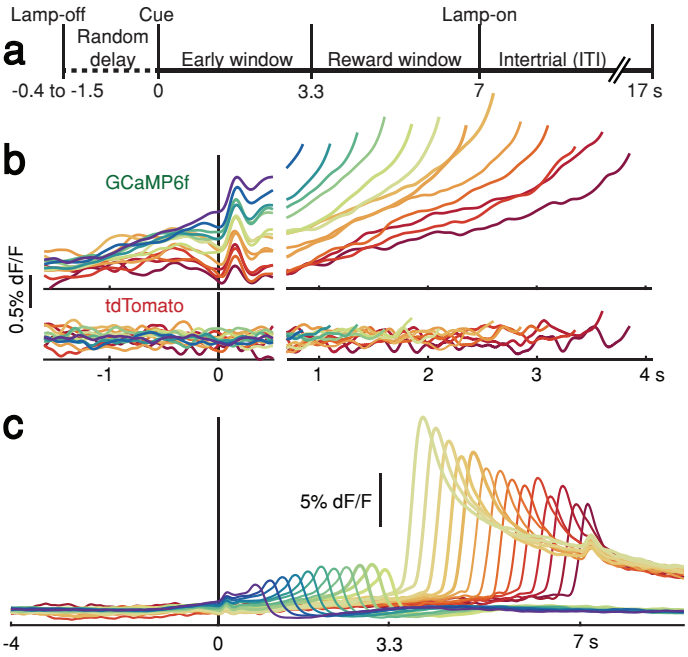

Figure 1 | Nigrostriatal dopaminergic signaling during a self-timed movement task. Figure adapted from Hamilos *et al.*, 2020[1] and used with permission. **a,** Schematic of self-timed movement task. **b,** Top**:** Average DAN GCaMP6f responses from 12 mice; Bottom: Responses of tdTomato, a non-activity-dependent fluorophore used to control for optical artifacts. The different colored traces correspond to averaged trial responses with different first-lick times (ranging from 1-4 s in 250 ms increments). Traces are plotted up to 150 ms before first-lick. Averaged traces are aligned relative to both start-timing cue onset (left of x-axis break) and first-lick (right of x-axis break); the break in the x-axis indicates the change in plot alignment. **c,** Cue-aligned average DAN GCaMP6f signals at lower gain show post-movement RPE-like signals. Movement onset occurs just before the peak response for each curve. Mice were rewarded for first-licks made later than 3.3 s, but were not rewarded for earlier first-licks.

A temporal-difference learning model of dynamic dopaminergic signaling

We were interested in understanding the origin of the dynamic dopaminergic signals we observed in our self-timed movement task and how they fit into the context of prior work on the dopamine system. A framework that has explained many disparate experimental results from the dopaminergic system is temporal difference (TD) learning with reward-prediction errors (RPE)[2,14]. In this framework, DAN activity is thought to reflect the moment-to-moment difference in the animal's expectation versus its perception of the value of its current state, where value is defined as the temporally-discounted expectation of total future reward. In classical trace-conditioning paradigms, DANs fire in transient bursts to unexpected rewards and reward-predicting cues, whereas they pause their firing when expected reward is omitted. Indeed, we observed RPE-like

signals in the cue-related transient, dips in activity after unrewarded first-licks, and surges in activity following rewarded first-licks (Figure 1c). Persistence of RPE-like signals in well-trained animals has been suggested to arise from the inherent imprecision in neural timing[10], which may reflect the animal's moment-to-moment uncertainty of its current state—i.e., its position in time–and, by extension to our task, uncertainty about its accuracy for a given self-timed lick[3]. Indeed, positive-going RPE-like signals were strongest for first-licks closest to the reward-boundary (3.3 s), presumably when the mouse's "confidence" of reward was lowest, consistent with the greatest RPE occurring when the mice were least certain of success (Figure 1c).

Whereas RPE-frameworks have explained *transient* bursts and pauses in DAN activity during trace conditioning and other types of learning experiments, DAN activity can also change more slowly[2,3]. For example, "ramping" signals build up over seconds during goal-directed navigation[15], bandit tasks in which animals must complete multiple goals to receive reward[16,17], and tasks with visual cues of proximity to reward[18]. It has been suggested that DANs could signal different information via slow changes in activity (e.g., motivation, ongoing value, vigor) compared to fast-timescale activity (e.g., post-hoc RPE signals for learning), and a number of proposals have suggested that DANs multiplex different kinds of information over different timescales and contexts[17,19].

However, recent models have proposed RPE-based explanations that may be able to reconcile these seemingly disparate dopamine signals[2,3,18]. While these models do not refute the possibility that DANs could encode other types of information (e.g., value, vigor, etc.), they are attractive for their parsimonious explanation of how fast time-scale phenomena and slowly-evolving ramps

could arise from the same underlying RPE-based calculation. In short, these models employ principles from TD learning to show how certain shapes of the value function (i.e., the assignment of values to the series of behavioral states comprising a task) can give rise to a *continuously changing* RPE, even in well-trained animals[2,3,18,20].

We were interested in whether an RPE-based framework could explain the results found in our self-timed movement task as well as results from other timing tasks[5]. To approach this question, we applied a key feature of TD learning algorithms to determine what an RPE-like signal would look like in different kinds of timing tasks. Specifically, we took advantage of the fact that *RPE is proportional to the derivative of the subjective value function under conditions of state uncertainty*[2,3], as is the case during timing tasks in which the animal must rely on its own internal representation of time to guide behavior[2].

Thus, if the value landscape for a given behavioral task is known, and if DAN activity encodes RPE, the RPE-based framework makes predictions about the expected shape of dynamic DAN activity during the task. In a recent study, similar applications of this principle predicted the ramping DAN signals that were observed in virtual reality (VR) tasks in which animals were moved passively through VR spaces, as well as when the animals passively viewed abstract, dynamic visual cues indicating proximity to reward[18], suggesting the ramping in our task could be explained from similar principles.

RPE-predictions for DAN responses during self-timed movement

In a simple TD learning model of self-timed movement, time may be modeled as a continuous set of states through which a Markov agent must traverse to receive reward[21] (Figure 2a). At each state transition (timestep), the agent must decide whether to move (lick) or to wait based on the probability of transitioning to a reward or failure state. If the agent is an optimal timer, its subjective approximation of its current state, τ, accurately tracks veridical time, t, and it will thus withhold movement until the first moment at which reward will be available in response to licking (3.3 s in our experiment).

The value landscape of this model can be understood intuitively. When the cue event occurs, a well-trained agent can expect an increased possibility of reward in the next few seconds; thus, at this moment, value increases. However, reward never occurs within the first 3.3 s of the standard timing task we implemented; thus, value at the cue is necessarily lower than value at 3.3 s. In fact, value will constantly increase as time approaches 3.3 s. Thus, as long as the agent withholds licks, the value landscape, V_t, during the first few seconds is a monotonically increasing, convex function[4] (Figure 2b). If the agent is an optimal timer, the subjective approximation of the value function, $\hat{V}_\tau$, matches the true value function, and $\hat{V}_\tau = V_t$.

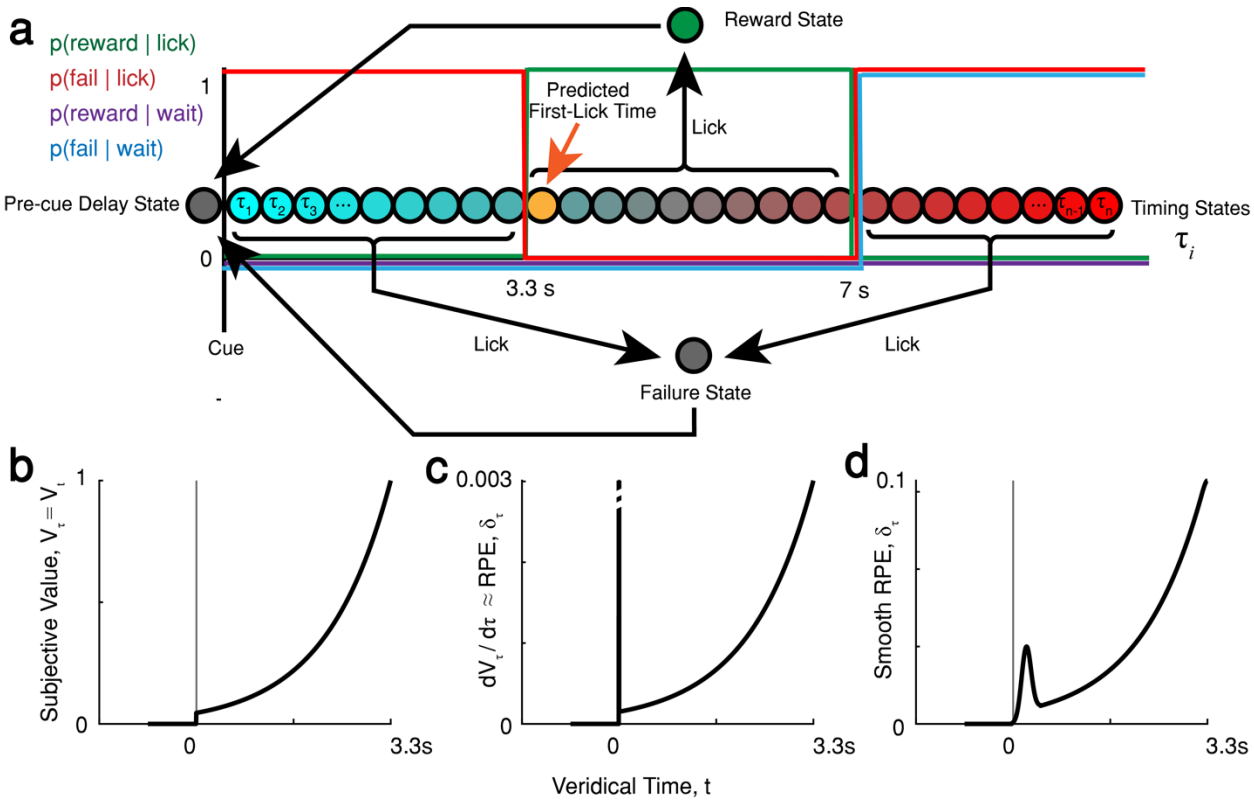

Figure 2 | Value and RPE Landscapes for an optimal timer predict DAN responses during the self-timed movement task. a, State space and probability of state transition for an optimal timer. Gold-shaded state is the first state from which reward is available, and thus is when the first-lick is predicted to occur. **b,** Estimated value function $\hat{V}_t$, where $\hat{V}_\tau \approx V_t$ for an optimal timer. An exponential value landscape is shown, consistent with prior literature[2]. However, any sufficiently convex function could be implemented with the same result[2-4]. The agent is expected to first-lick at the peak of the trajectory. **c,** RPE function for an optimal timer, estimated as $\delta_\tau \approx \hat{V}'_\tau$, the derivative of the subjective value function. Y-axis scaled to show ramp. **d,** Predicted DAN GCaMP6f signals for an optimal timer. The RPE function was smoothed with a gaussian kernel spanning *ca.* 10% of the interval to approximate GCaMP6f off-dynamics.

However, we assume that, because the timer does not have access to the true state identity, t, it is never certain of its subjective approximation of its state, τ. Under conditions of state uncertainty, RPE is approximately the derivative of the subjective value function[2,18], $\delta_\tau \approx \hat{V}'_\tau$, where δ_τ is RPE at subjective time τ, and $\hat{V}'_\tau$ is the time-derivative of the subjective value function. Thus, the shape of the RPE function, δ_τ is also quite simple: a transient increase at the cue followed by a

slowly-evolving ramp (Figure 2c). If the RPE function is measured by a calcium indicator such as GCaMP6f, the binding kinetics of the indicator would tend to blur the RPE function, which we approximated by smoothing (Figure 2d).

The modeled RPE function mirrors the shape of the dynamics observed in DAN signals: a cue-related transient followed by a slow ramp up to the time of first-lick. However, unlike the optimal timer in this model, mice, like humans, exhibit suboptimal timing behavior with variability proportional to the duration of the timed interval[10]. It has been proposed that this variability in timing results from imprecision in an internal clock, referred to classically as the internal "pacemaker[22]". When the pacemaker is fast, self-timed movements occur relatively early, whereas when the pacemaker is slow, later movements occur. These changes in the pacemaker rate would correspond to the mouse traversing the set of subjective states, τ, at different rates than the passage of veridical time, t (Figure 3a), resulting in relative *compression* and *stretching*, respectively, in the subjective value function, $\hat{V}_\tau$ (Figure 3b), with corresponding compression/stretching of the RPE function (Figure 3c).

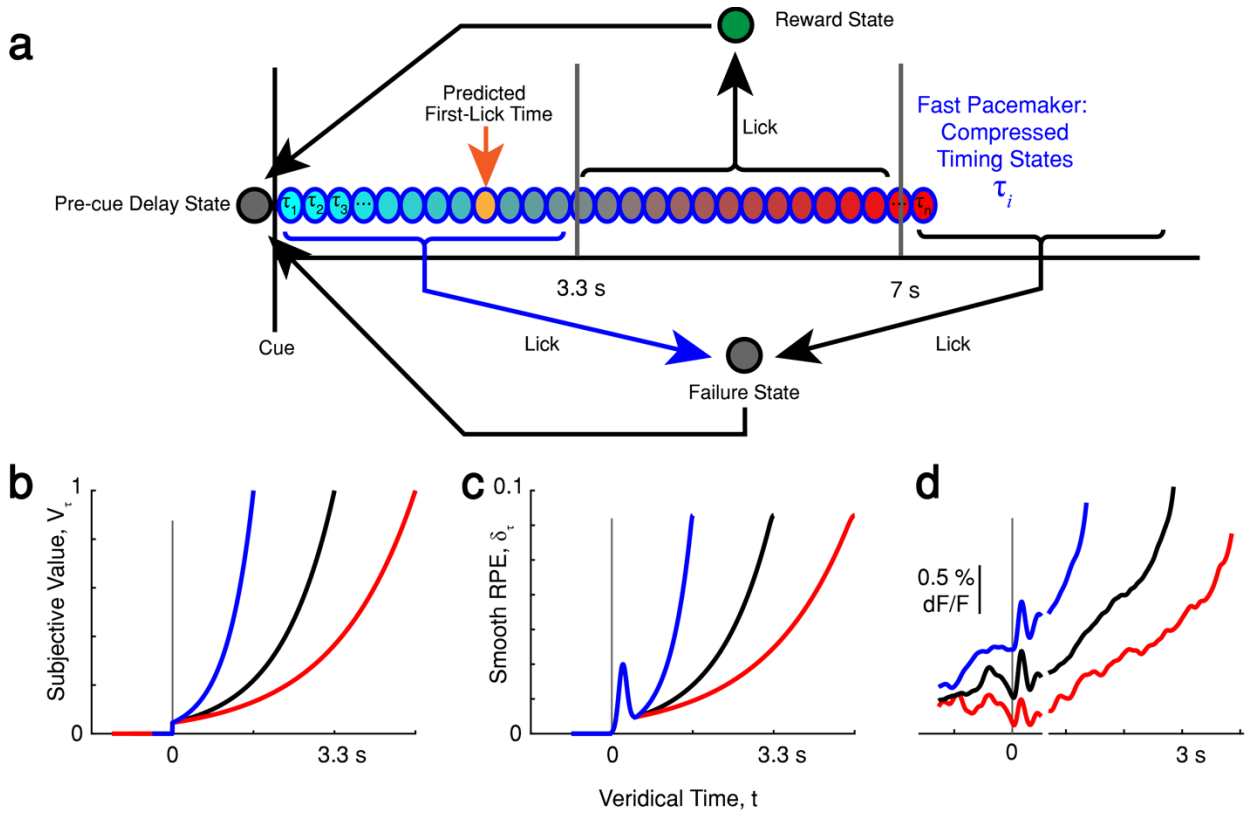

Figure 3 | Compressed and stretched Value and RPE Landscapes for a sub-optimal timer predict dynamic DAN responses during the self-timed movement task, but do not capture baseline offsets. a, Simple state space of self-timed movement task for a suboptimal timer with a fast pacemaker. The fast pacemaker "compresses" state space[3,21], resulting in traversal of the timing states faster than veridical time. The mouse can only make a decision based on which state it believes itself in; thus first-lick is expected to occur early (gold-shaded state). **b,** A compressed subjective value function ($\hat{V}_\tau$, blue) reflects relatively fast traversal through the value landscape compared with that of veridical time (V_t, black). Conversely, stretched $\hat{V}_\tau$ (red) reflects slow traversal, consistent with a slow pacemaker. The animal is expected to lick at the peak of the trajectory. **c,** Smoothed estimated RPE function ($\hat{V}'_\tau \approx \delta_\tau$). Compression/stretching of the value function produces ramping dynamics similar to those observed in DANs (**d**) and striatal dopamine[1]. However, this model alone does not explain the more tonic baseline offsets that were anti-correlated with upcoming movement time (**d** and Figure 1b). **d,** Average DAN GCaMP6f signals (12 mice, 3 timepoints replotted from Figure 1b, plotted up to 150 ms before first-lick). Break in x-axis as in Figure 1b.

Strikingly, as this simple RPE-based model predicts, DAN signals observed during our self-timed movement task show different ramping dynamics depending on when the animal actually moved (Figure 3d), consistent with compression/stretching of the subjective value and RPE functions.

When the animal moved relatively early (perhaps corresponding to a fast pacemaker), DAN ramping unfolded with a steeper slope, as if the ramping period were *compressed*. Conversely, when the animal moved late (perhaps corresponding to a slow pacemaker), DAN ramping unfolded with a shallower slope, as if the ramping interval were *stretched*. The idea of compression/stretching of DAN ramps was supported by our encoding model[1], for which we needed to add a timing-dependent "stretch factor" to best capture the variance in GCaMP6f signals during the timed interval. Together, these observations could be explained by DANs encoding an RPE-like signal related to the animal's "belief" of its position in objective time, τ, as derived from its position along the subjective value trajectory during the timing interval of the task.

In fact, a recent model described how a timing mechanism instantiated by the nigrostriatal system could lead to (the well-known) variability in self-timed intervals by stretching or compressing of subjective value trajectories[3]. The model posits that dopamine modulates the pacemaker rate (consistent with pharmacological and lesion studies), with increased dopamine availability (or efficacy) speeding the pacemaker, and decreased dopamine slowing the pacemaker[6-8,11-13]. In turn, the pacemaker controls the encoding of subjective time, and thus the steepness of the value function with respect to objective, veridical time. It follows that variation in dopamine availability would compress or stretch the value landscape to varying degrees from trial-to-trial. This model is consistent with our findings of variable ramping slope in DANs signals from trial-to-trial. It is also consistent with neural recordings from striatal spiny projection neurons and parietal cortical neurons during similar self-timed movement tasks, for which temporal sequences of striatal and cortical firing during timing were compressed for early movements and stretched for late movements[23,24].

While the RPE-based view of DAN activity captures the dynamic DAN signals we observed, our simple RPE model alone does not capture the *baseline offsets* in DAN signals that were predictive of movement timing even after controlling for previous trial outcome and ongoing nuisance movements[1]. More complex RPE-based explanations for these *tonic* offsets in DAN signals could be imagined with further assumptions (e.g., states like the pre-cue delay could also contain timing states that create offsets before the trial begins, etc.), but a parsimonious explanation for how and why these offsets emerge requires further investigation. Mohebi *et al*. recently showed baseline differences in the amount of dopamine in the nucleus accumbens core that were correlated with the recent history of reward rate: higher recent reward rates were related to higher tonic dopamine[17]. However, in our task, animals tended to move later toward the end of sessions, resulting in periods of relatively high reward rate when the average tonic baseline signal was *lower* (baseline preceding rewarded trials—by definition, later movements—was systematically lower in our task, Figure 1b-c), suggesting a more complex relationship between tonic DAN activity and reward rate in our task. While the origin of offsets in DAN signals remains unclear, these offsets were nonetheless inversely related to the first-lick time, and thus directly related to the (inferred) pacemaker rate, consistent with pharmacological and lesion studies positing a positive correlation between dopamine availability and pacemaker rate[3,6-8,11-13].

Ramping signals in our photometry experiments were measured from a population of DANs. An important future question is whether ramps are also present at the level of individual neurons, or rather represent a progressive recruitment of individual neurons, or some combination of both. Prior studies have reported ramping signals in individual neurons during tasks with visual feedback

263 of distance to reward[18], whereas others have observed decoupling between DAN firing rates and
264 downstream dopamine release[17], making it unclear whether electrophysiology would be capable
265 of addressing this question. Observation of individual neurons expressing calcium indicators with
266 GRIN-lens equipped endoscopes may be better suited to this question.
267
268 ***RPE-based predictions for DAN responses during a temporal bisection task***
269 Whereas DAN signals during our self-timed movement task were consistent with classic
270 observations of the influence of dopamine on the speed of the pacemaker, a recent study employing
271 a different timing task found more complex DAN dynamics during timing. Soares *et al*. recorded
272 SNc DAN GCaMP6f signals with fiber photometry as mice executed a classic temporal bisection
273 perceptual task[5] (Figure 4a). Trials began when mice entered a nose-poke port and received an
274 auditory start-timing cue. Mice had to remain in the port throughout a variable timing interval,
275 which was terminated with a stop-timing auditory cue. Mice then reported whether the interval
276 was shorter or longer than a criterion time (1.5 s) by choosing a left or right nose-poke port
277 corresponding to a "long" or "short" judgment. Mice were trained to categorize intervals spanning
278 0.6-2.4 s. As expected, trials with more extreme intervals were easier for the mice, whereas trials
279 with intervals closer to the 1.5 s criterion time elicited chance performance (Figure 4b).
280

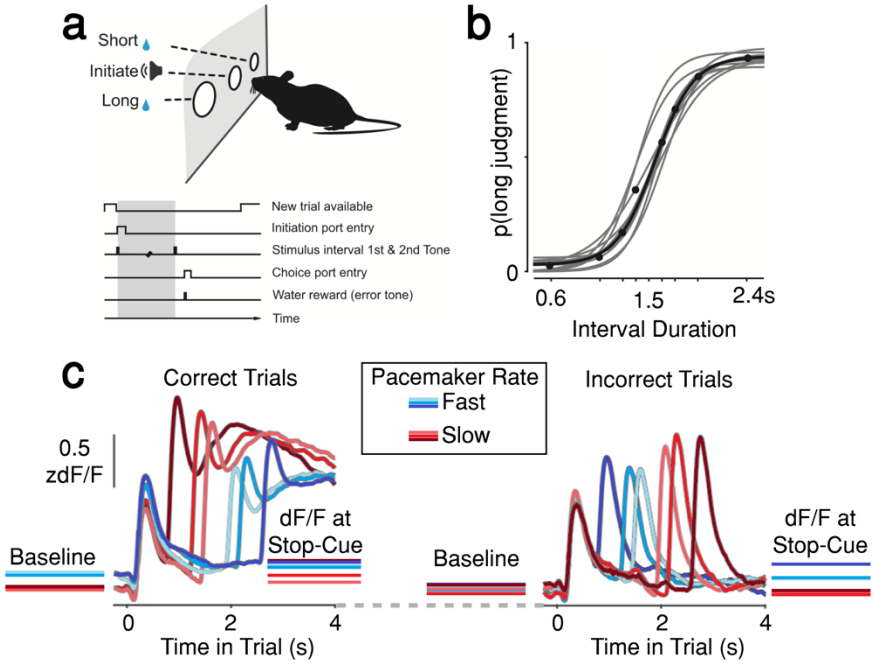

Figure 4 | A temporal bisection task shows relatively high DAN signals during the timing interval when the inferred pacemaker rate is relatively fast. Figures adapted from Soares *et al.*, 2016[5] with permission of authors and AAAS. **a,** Task schematic. **b,** Psychometric curve for timing intervals of different duration. Criterion time: 1.5 s. **c,** Start-timing cue-aligned average SNc DAN GCaMP6f signals. Second peak occurs just after the stop-timing cue (intervals: 0.6, 1.05, 1.26, 1.74, 1.95, 2.4 s). Figure recolored to indicate average inferred pacemaker rate. Red: slow; blue: fast. Note: colors intended to indicate category of clock speed, *not* relative pacemaker speed within category. Relative dF/F amplitude during baseline and immediately prior to stop-timing cue shown left and right. dF/F amplitudes during timing are higher when the inferred pacemaker rate is fast. Left: Correct trials. Right: Incorrect trials show the same dF/F relationship with pacemaker rate.

DANs exhibited complex dynamics during the bisection task, starting with a sharp transient after the start-timing cue and ending with a second transient after the stop-timing cue (Figure 4c). Between the start-timing and stop-timing cues, DAN signals exhibited a U-shape with increasing time, which was visible for trials with longer intervals but was truncated prematurely for the shorter intervals. The authors focused their analyses on the transient occurring *after the stop-timing* cue. Short judgments (suggesting a slow pacemaker) were accompanied by relatively high-amplitude transients after the stop-cue, whereas long judgments (suggesting a fast pacemaker) showed

301 relatively low-amplitude transients. These results seemed to suggest that relatively *high* DAN
302 activity reflected a *slow* pacemaker, the opposite of what is expected based on the bulk of
303 pharmacological and lesion studies[3], as well as the trend we observed during our self-timed
304 movement task.
305
306 This surprising finding could be a unique feature of the bisection task. Unlike self-timed
307 movements, in which animals directly report elapsed time with a movement, the temporal bisection
308 task requires an additional computational step, in which the timed interval must be categorized as
309 "long" or "short." However, prior pharmacological studies employing the bisection task found
310 results consistent with the classic view that higher dopamine availability is associated with a faster
311 pacemaker[3,25]—opposite the interpretation of Soares *et al*., but consistent with the findings of our
312 self-timed movement task.
313
314 The discrepancy between our results and those found by Soares *et al*. could perhaps be traced to
315 differences in the way DAN signals were analyzed. We focused our attention on DAN signals
316 unfolding *during timing* in our self-timed movement task, whereas these signals were not explored
317 by Soares *et al*. We thus asked two questions: 1. What correlations exist between DAN signals and
318 pacemaker rate in the bisection task *before* the timing interval? And, 2. What correlations exist
319 *during* the timing interval itself?
320
321 Before addressing these questions, we note that the relationship between pacemaker and bisection
322 judgment is not as straightforward as in self-timed movement, and thus we recolored Figure 4c to
323 clarify this, employing the following intuition: For a trial to be correct in the bisection task, on

average, the pacemaker must be either accurate or "conservatively inaccurate." In other words, a correct "short" judgment requires either accurate timing or a *slow* pacemaker (Figure 4c, red curves). Conversely, a correct "long" judgment requires either accurate timing or a *fast* pacemaker (Figure 4c, blue curves).

When we considered DAN signals *before* the timing interval for correct trials in the Soares *et al*. study (Figure 4c, left), we noticed what appears to be two strata of signal levels. Trials with "long" judgments (fast pacemaker on average) had relatively high baseline signals, whereas trials with "short" judgments (slow pacemaker on average) had lower baseline signals, consistent with the relationship between baseline offsets and pacemaker rate that we observed in our self-timed movement task. As in our task, these baseline offsets remained present during the timing interval, resulting in the same stratification of dF/F signals immediately prior to the stop-timing cue (except for the very shortest interval, 0.6 s, which overlaps decaying GCaMP6f signals related to the start-timing cue, likely causing an artifactual inflation of the signal just prior to the stop-cue due to the off-kinetics of the calcium indicator or kinetics of calcium clearance more generally). Thus, it generally appears that DAN activity was *higher* on trials with fast pacemaker rates, both during and before the interval in which the animal was actually timing. Intriguingly, *incorrect* trials (to the right in Figure 4c) showed a relative convergence of the baseline signals preceding the start-cue, but then signals diverged during the timing interval, resulting in relatively *high* signals just before the stop-cue for incorrect "long" choices (i.e., a fast pacemaker, blue), but relatively low signals just before the stop-cue for incorrect "short" choices (i.e., a slow pacemaker, red). This is consistent with the patterns observed on correct trials. Interpreted thusly, the Soares *et al*. result is consistent both with our results and with classic pharmacological studies relating higher/lower

347 dopamine availability to faster/slower pacemaker rates, respectively. Soares *et al*. presented their
348 subsequent analyses with these baseline differences normalized-out in some way (Figure 3 of
349 Soares *et al*.). It is possible that this "zeroing out" of the baseline offset may have hindered efforts
350 to detect consistent effects during the timing interval due to reordering of the traces.
351
352 Because baseline offsets in the bisection task appear similar to those in our self-timed movement
353 task, we asked whether dynamic DAN signals in the bisection task could similarly be explained
354 by the task's RPE landscape. In their investigation of the stop-timing cue-related transient, Soares
355 *et al*. showed that its amplitude is well-explained by a combination of temporal surprise and
356 behavioral performance, and we applied these parameters to derive a value landscape consistent
357 with their bisection task.
358
359 The inferred value landscape of the bisection task for an optimal agent was built from a few
360 assumptions (Figure 5a):
361
362 1. As in our self-timed movement task, value increases immediately at the start-cue and
363 continues to rise toward the time of expected potential reward delivery.
364
365 2. Because the longest interval is 2.4 s, the time until potential reward is known to be no more
366 than ~3 s (including the time to report judgment). However, due to temporal uncertainty
367 and the fact that a false start (leaving the port before the stop-timing cue) results in an error
368 and loss of reward, there is a second jump in the value function at the time of the stop-cue

when the feedback of the tone reorients the value function and indicates the opportunity to collect reward within a few hundred milliseconds.

3. Because value is temporally discounted at the start-cue by the possibility of the longest-possible interval, any stop-cue occurring before 2.4 s results in a sudden "teleportation" through the value landscape to the final limb of the task that occurs just before the judgment and ascertainment of trial outcome, similar to the jump in the value function in a recently-reported, virtual reality, spatial teleportation task[18]. Thus, assuming the value function trends upwards steadily, the amplitude of RPE-related transients following the stop-cue would *decrease* as the interval duration increases, because the sudden jump in the value function becomes progressively smaller.

4. To capture aspects related to behavioral performance, we additionally included contours in the value function during the timing interval to reflect the probability of a correct choice for intervals of different lengths. Specifically, a relative minimum in the value function occurs near 1.5 s, when predicted performance is worst. However, a stop-timing tone near the criterion time also results in a smaller jump in the value function because the probability of a correct decision is also lower. Thus, the increase in value at the moment of decision was adjusted by the probability of a correct choice.

5. As in the simple RPE-model of our self-timed movement task, we modeled changes in pacemaker rate as compression/stretching of the subjective value landscape with respect to veridical time.

6. The agent traverses timing states during the timing interval, similar to the timing states in the self-timed movement task, but unlike our task, the bisection task does not require the agent to decide when to move. We assume the need to make a timed movement imposes a need for the agent to be relatively certain of its subjective timing state, τ, to make a decision, even though it is uncertain of its true state, t. The bisection task, on the other hand, is more similar to classical conditioning tasks in which the timing interval is not in the agent's control, and thus subjective state uncertainty increases with the distance from the last state-informative cue[3]. Thus, we took into account temporal blurring of the subjective state function, which would tend to reduce the convexity of the subjective value function and reduce the amplitude of ramping during the timing interval[3]. However, adding temporal blurring does not substantially change the fit-shape in our simplified model, and versions with or without blurring can reproduce the shape of the dynamic DAN signals.

Together, we arrived at a model of the RPE landscape for each of the six tested interval durations (Figure 5b,c). Importantly, this simple RPE-based model accurately captures the relative categorical amplitudes of the stop-timing cue-related transients, as follows: If the instantaneous DAN activity at the time of the stop-timing cue is relatively high, this would indicate that the animal is further along in the subjective value trajectory, resulting in 1) a *long* judgment, and 2) a relatively *smaller* RPE transient, because the underlying subjective value was higher at that moment. Conversely, if instantaneous DAN activity is relatively low at the stop-timing cue, this would indicate that the animal is not very far along the subjective value trajectory, leading to 1) a

short judgement and 2) a relatively *larger* stop-cue-related RPE transient, because the underlying subjective value was relatively low just before the stop-cue.

Now consider a particular (objective) time interval near the criterion time, for which the animal makes a mix of "long" and "short" choices (e.g., 1.74 s; Figure 4b). Soares *et al*. found that the amplitude of the stop-timing cue-related GCaMP6f transient tended to be bigger when the animal incorrectly made short choices, and this was taken as evidence that elevated DAN activity *slows* the internal clock. However, our model predicts that the size of the stop-cue-related transient will be inversely related to the amplitude of the underlying subjective value at that point, and thus inversely related to elapsed *subjective* time. It thus follows that if subjective time is more advanced on a given trial (i.e., faster pacemaker), the animal would tend to choose the long judgment on that trial, and the stop-timing RPE transient would be *smaller*. Conversely, if subjective time is less advanced on a trial (i.e., slower pacemaker), the animal would tend to choose the short judgment, and the stop-timing RPE transient would be *larger*.

Our RPE model accurately predicts the results of Soares *et al*.; however, our model holds that elevated DAN activity *speeds* the internal clock, consistent with most pharmacological studies but *opposite* the interpretation of Soares *et al*. Thus, our RPE-based model suggests a parsimonious explanation for DAN activity in both the self-timed movement and temporal bisection paradigms, with (1) relatively high DAN activity corresponding to a fast pacemaker; manifesting in (2) compression of the value landscape; thereby leading to (3) early movements (in the self-timed movement task) or long judgments (in the temporal bisection task).

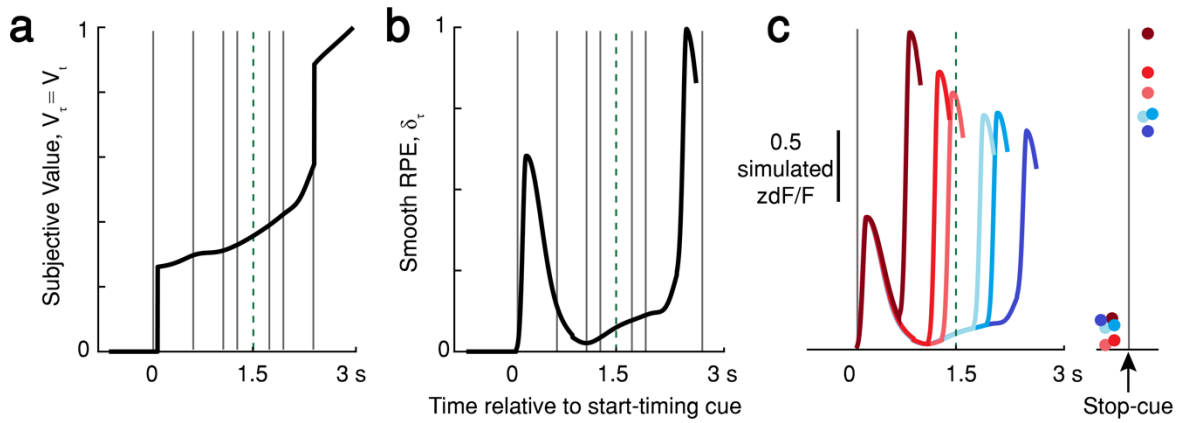

Figure 5 | Subjective Value and RPE Landscapes for the temporal bisection task predict dynamic DAN responses during the temporal bisection task, but do not capture baseline offsets. a, Estimated value function $\hat{V}_t$, where $\hat{V}_\tau \approx V_t$ for an optimal timer on a 2.4 s trial. Grey lines: test interval times. Green dashed line: criterion time (1.5 s). Value increases approaching the time when reward is available, increasing abruptly at the start- and stop-timing cues (0 and 2.4 s). **b,** Smoothed RPE function for an optimal timer, estimated as $\delta_\tau \approx \hat{V}'_\tau$, the derivative of the subjective value function. The RPE function was smoothed with an asymmetrical gaussian kernel spanning *ca*. 28% of the interval to approximate GCaMP6f off-dynamics. **c,** Predicted DAN GCaMP6f signals for an optimal timer for the six test interval times. Traces truncated before reward collection for clarity. Colors indicate conservative pacemaker speed for a correct judgment (red: slow, blue: fast). Right: relative simulated dF/F amplitude just before the stop-timing cue and subsequent peak response. *Amplitude just before the stop-timing cue is directly proportional to pacemaker speed; peak amplitude after the stop-timing cue is inversely proportional to pacemaker speed.*

Limitations of the RPE-based model

The simple RPE-based models presented here explain dynamic DAN signals in both the bisection task and our self-timed movement task, but they do not explain the origin of baseline offsets. Mohebi *et al*.[17] recently-proposed that baseline offsets in ventral striatal dopamine levels could reflect the average recent reward rate, but we found that offset amplitude in DAN signals is at least partially independent of recent trial history during the self-timed movement task. It is possible that baseline variation arises from slow, random fluctuations in DAN activity, but further work is needed to explore the origins of these signals.

A second issue is the impact of optogenetic DAN activation and suppression on the rate of the pacemaker. In our self-timed movement task, DAN activation promoted early movements, consistent with increasing the pacemaker rate, whereas suppression promoted late movements, consistent with slowing the pacemaker rate[1]. However, Soares *et al*. reported an opposite effect for optogenetic manipulation during the bisection task, at least for DAN activation[5].

This difference between the tasks could be reconciled by a recent theoretical model proposed by Mikhael and Gershman to explain the behavior of the pacemaker in a wide range of classical conditioning and timing studies[3]. Their model shows that the pacemaker rate is expected to be updated at the time of reinforcement by a Hebbian-like, bidirectional learning rule. If reward occurs exactly at the expected time, there is no update in the pacemaker rate. However, if reinforcement occurs before the expected time, this is interpreted as feedback that the pacemaker was running too slowly; thus, the update rule increases the pacemaker rate leading to expectation of reward at an earlier time on the next trial. Conversely, if reinforcement occurs after it was expected, this is interpreted as feedback indicating an overly fast pacemaker, resulting in an update that slows the pacemaker rate and creates the expectation of a later reward on the next trial. The same principles apply to ongoing RPE during timing tasks.

In our self-timed movement task, we activated or inhibited DANs only *up to* the time of first-lick, which Mikhael and Gershman's model predicts will produce an effect on the pacemaker rate consistent with the sign of the manipulation (activate: increase, inhibit: decrease). However, Soares *et al*. continued optical stimulation *past* the end of the timing interval, until the end of the trial. When Mikhael and Gershman modeled stimulation in the Soares *et al*. task, they found that simulated DAN activation increased the pacemaker rate during the timing interval, but the

continuing stimulation after the stop-timing cue rapidly counteracted this effect, resulting in *slowing* of the modeled pacemaker between the stop-cue and the judgment, leading to an effect on pacemaker rate *inconsistent* with the sign of the manipulation, as observed in Soares *et al*. If this model is correct, the effect of stimulation on the animal's judgment in the Soares *et al*. task may have arisen due to continued manipulation of DAN activity *after* the timing interval had ended. A "retrospective" effect of this sort might seem counterintuitive, but such retrospective effects have long been observed in perceptual studies, in which recall of sensory stimuli can be enhanced by additional sensory cues presented shortly after stimulus offset[26,27], suggesting that sensory events are "buffered" briefly and can be altered by neural activity occurring between the sensory event and the perceptual decision. It is possible that a similar process could occur in the bisection task if DAN stimulation extends past the timing interval, although this is speculative. More work is needed to reconcile the optogenetic results in the self-timed movement and bisection tasks. To start, it would be informative to repeat the optogenetic experiments in the bisection task with optical stimulation limited to the period of the timed intervals only.

REFERENCES

1. Hamilos, A.E. *et al*. Dynamic dopaminergic activity controls the timing of self-timed movement. Preprint at: https://doi.org/10.1101/2020.05.13.094904 (2020).
2. Mikhael, J. G., Kim, H. R., Uchida, N., & Gershman, S. J. Ramping and State Uncertainty in the Dopamine Signal. Preprint at https://www.biorxiv.org/content/10.1101/805366v1 (2019).
3. Mikhael, J. G. & Gershman, S. J. Adapting the flow of time with dopamine. *J. Neurophysiol*., **121**, 1748–1760 (2019).
4. Gershman, S. J. Dopamine Ramps Are a Consequence of Reward Prediction Errors. *Neural Comp*. **26**, 467–471 (2014).
5. Soares, S., Atallah, B. & Paton, J. Midbrain dopamine neurons control judgment of time. *Science* **354,** 1273–1277 (2016)
6. Dews, P. B. & Morse, W. H. Some observations on an operant in human subjects and its modification by dextro amphetamine. *J. Exp. Anal. Behav*. **1,** 359–364 (1958).
7. Schuster, C. & Zimmerman, J. Timing behavior during prolonged treatment with dl-amphetamine. *J. Exp. Anal. Behav*. **4,** 327–330 (1961).
8. Meck, W. H. Affinity for the dopamine D2 receptor predicts neuroleptic potency in decreasing the speed of an internal clock. *Pharmacol. Biochem. Behav*. **25,** 1185–1189 (1986).
9. Malapani, C., Rakitin, B. C., Levy, R., Meck, W. H., Deweer, B., Dubois, B., & Gibbon, J. (1998). Coupled Temporal Memories in Parkinson's Disease: A Dopamine-Related Dysfunction. J. Cog. Neuro., *3*(May), 316–331.

10. Raitin, B. C., Penney, T. B., Gibbon, J., Hinton, S. C., & Meek, W. H. Scalar Expectancy Theory and Peak-Interval Timing in Humans. *Journal of Experimental Psychology: Animal Behavior Processes* **24**, 15–33 (1998).
11. Lutig, C. & Meck, W. H. Chronic treatment with haloperidol induces deficits in working memory and feedback effects of interval timing. *Brain Cogn*. **58,** 9–16 (2005).
12. Meck, W. H. Neuroanatomical localization of an internal clock: A functional link between mesolimbic, nigrostriatal, and mesocortical dopaminergic systems. *Brain Res*. **1109,** 93–107 (2006).
13. Merchant, H., Harrington, D. L. & Meck, W. H. Neural Basis of the Perception and Estimation of Time. *Annu*. *Rev*. *Neurosci* 36, 313–36 (2013).
14. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
15. Howe, M. W., Tierney, P. L., Sandberg, S. G., Phillips, P. E. M. & Graybiel, A. M. Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature* **500**, 575–579 (2013).
16. Hamid, A. A. et al. Mesolimbic dopamine signals the value of work. *Nat*. *Neurosci*. **19**, 117–126 (2016).
17. Mohebi, A et al. Dissociable dopamine dynamics for learning and motivation. *Nature* **570**, 65–70 (2019).
18. Kim, H. R. et al. A unified framework for dopamine signals across timescales. Preprint at https://www.biorxiv.org/content/10.1101/803437v1 (2019).

19. Engelhard, B et al. Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature* **570**, 509–513 (2019).

20. Starkweather, C. K., Babayan, B. M., Uchida, N., & Gershman, S. J. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat. Neuro.* **20**, 581–589 (2017).

21. Sutton, R. S., & Barto, A. G. *Reinforcement Learning*. MIT Press, Cambridge, (2000).

22. Gibbon, J., Malapani, C., Dale, C. L., & Gallistel, C. Toward a neurobiology of temporal cognition: advances and challenges. *Curr. Opin. Neurobiol.* **7**, 170–184 (1997).

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

24. Maimon, G. & Assad, J. A. A cognitive signal for the proactive timing of action in macaque LIP. *Nat. Neuro.* ***9***, 948–955 (2006).

25. Morgan, L., Killeen, P.R., Fetterman, J.G. Changing rates of reinforcement perturbs the flow of time. *Behav. Processes* **30**, 259–271 (1993).

26. Gegenfurtner, K. R., & Sperling, G. Information Transfer in Iconic Memory Experiments. *J. Exp. Psych: Human Perception and Performance* 19, 845–866 (1993).

27. Herrington, T. M., & Assad, J. A. Neural activity in the middle temporal area and lateral intraparietal area during endogenously cued shifts of attention. *J. Neurosci.* **29**, 14160–14176 (2009).

ACKNOWLEDGEMENTS

We thank J.G. Mikhael and S.J. Gershman for discussions on temporal difference learning models and analytical methods; The work was supported by NIH grants UF-NS109177 and U19-NS113201, and NIH core grant EY-12196. A.E.H. was supported by a Harvard Lefler Predoctoral Fellowship and a Harvard Quan Predoctoral Fellowship.

AUTHOR CONTRIBUTIONS

A.E.H. and J.A.A conceived the project. A.E.H. performed all experiments and implemented the computational models. A.E.H. and J.A.A. analyzed the data and wrote the paper.

DECLARATION OF INTERESTS

The authors have no relevant interests to declare.

CODE AVAILABILITY

All custom behavioral software and analysis tools are available at https://github.com/harvardschoolofmouse.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.