Article

Neuron

Temporal Specificity of Reward Prediction Errors Signaled by Putative Dopamine Neurons in Rat VTA Depends on Ventral Striatum

Highlights

- Dopamine neurons signal errors in reward prediction due to shifts in timing or number
- Without ventral striatum, dopamine neurons signal errors in reward number normally
- However, the same neurons fail to signal errors in reward timing.
- These data indicate that these elements of the predictions are neurally dissociable.

Authors

Yuji K. Takahashi, Angela J. Langdon, Yael Niv, Geoffrey Schoenbaum

Correspondence

yuji.takahashi@nih.gov (Y.K.T.), geoffrey.schoenbaum@nih.gov (G.S.)

In Brief

Using single-unit recording and computational modeling, Takahashi et al. show that predictions about reward timing and quantity are dissociable and that dopaminergic error signals rely on the ventral striatum for the former but not the latter.

Takahashi et al., 2016, Neuron *91*, 182–193
July 6, 2016 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.neuron.2016.05.015

CellPress

Neuron
Article

Temporal Specificity of Reward Prediction Errors Signaled by Putative Dopamine Neurons in Rat VTA Depends on Ventral Striatum

Yuji K. Takahashi,[1,5,*] Angela J. Langdon,[2,5] Yael Niv,[2] and Geoffrey Schoenbaum[1,3,4,*]
[1]NIDA Intramural Research Program, Baltimore, MD 21224, USA
[2]Department of Psychology and Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA
[3]Departments of Anatomy & Neurobiology and Psychiatry, University of Maryland School of Medicine, Baltimore, MD 21201, USA
[4]Solomon H. Snyder Department of Neuroscience, The Johns Hopkins University, Baltimore, MD 21287, USA
[5]Co-first author
*Correspondence: yuji.takahashi@nih.gov (Y.K.T.), geoffrey.schoenbaum@nih.gov (G.S.)
http://dx.doi.org/10.1016/j.neuron.2016.05.015

SUMMARY

Dopamine neurons signal reward prediction errors. This requires accurate reward predictions. It has been suggested that the ventral striatum provides these predictions. Here we tested this hypothesis by recording from putative dopamine neurons in the VTA of rats performing a task in which prediction errors were induced by shifting reward timing or number. In controls, the neurons exhibited error signals in response to both manipulations. However, dopamine neurons in rats with ipsilateral ventral striatal lesions exhibited errors only to changes in number and failed to respond to changes in timing of reward. These results, supported by computational modeling, indicate that predictions about the temporal specificity and the number of expected reward are dissociable and that dopaminergic prediction-error signals rely on the ventral striatum for the former but not the latter.

INTRODUCTION

Reward prediction errors are famously signaled, at least in primates and rodents, by midbrain dopamine neurons (Barto, 1995; Mirenowicz and Schultz, 1994; Montague et al., 1996; Schultz et al., 1997). Key to signaling reward prediction errors are reward predictions (Bush and Mosteller, 1951; Rescorla and Wagner, 1972; Sutton and Barto, 1998). Theoretical and experimental work has suggested that the ventral striatum (VS) is an important source of these predictions, particularly to dopamine neurons in the ventral tegmental area (VTA) (Daw et al., 2005, 2006; Joel et al., 2002; O'Doherty et al., 2003, 2004; Seymour et al., 2004; Willuhn et al., 2012). Here we tested this hypothesis by recording the activity of putative dopaminergic neurons in the VTA of rats performing a task in which positive and negative prediction errors were induced by shifting either the timing or the number of expected reward. We found that dopamine neurons recorded in sham-lesioned controls exhibited prediction error signals in response to both manipulations. By contrast, dopamine neurons in rats with ipsilateral VS lesions exhibited prediction error signals only in response to changes in the number of reward; these neurons failed to respond to changes in reward timing. Computational modeling of these data, using a framework that separates learning about reward timing from learning about reward number (Daw et al., 2006), showed that this pattern of results could be obtained by degrading the model's ability to learn precise timing, while leaving all other aspects of the model intact. Contrary to proposals that VS might be required for all aspects of reward prediction (Joel et al., 2002; O'Doherty et al., 2003; Willuhn et al., 2012), these results suggest that the VS is critical for endowing reward predictions with temporal specificity.

RESULTS

We recorded single-unit activity in the VTA of rats with ipsilateral sham (n = 9) or neurotoxic (n = 7) lesions of VS (see Figures S1A and S1B for recording locations). Recordings were made in rats with ipsilateral lesions to avoid confounding any neural effects of lesions with behavioral changes (Burton et al., 2014). Lesions targeted the VS core, resulting in visible loss of neurons in 57% (35%–75%) of this region across subjects (Figure S1F). Neurons in the lesioned area are known to fire to reward-predictive cues (Bissonette et al., 2013; O'Doherty et al., 2003, 2004; Oleson et al., 2012; Roesch et al., 2009) and send output to VTA (Bocklisch et al., 2013; Grace and Bunney, 1985; Groenewegen et al., 1990; Mogenson et al., 1980; Voorn et al., 2004; Watabe-Uchida et al., 2012; Xia et al., 2011), supporting the proposal that this part of VS may influence dopaminergic prediction error signaling as proposed in neural instantiations of temporal difference reinforcement learning (TDRL) models (Daw et al., 2006; Joel et al., 2002).

Neurons were recorded in rats performing an odor-guided choice task used previously to characterize signaling of reward predictions and reward prediction errors (Roesch et al., 2007; Takahashi et al., 2009, 2011). On each trial, rats sampled one of three different odor cues at a central port and then responded at one of two adjacent wells (Figure 1A). One odor signaled the availability of sucrose reward only in the left well (forced left), a

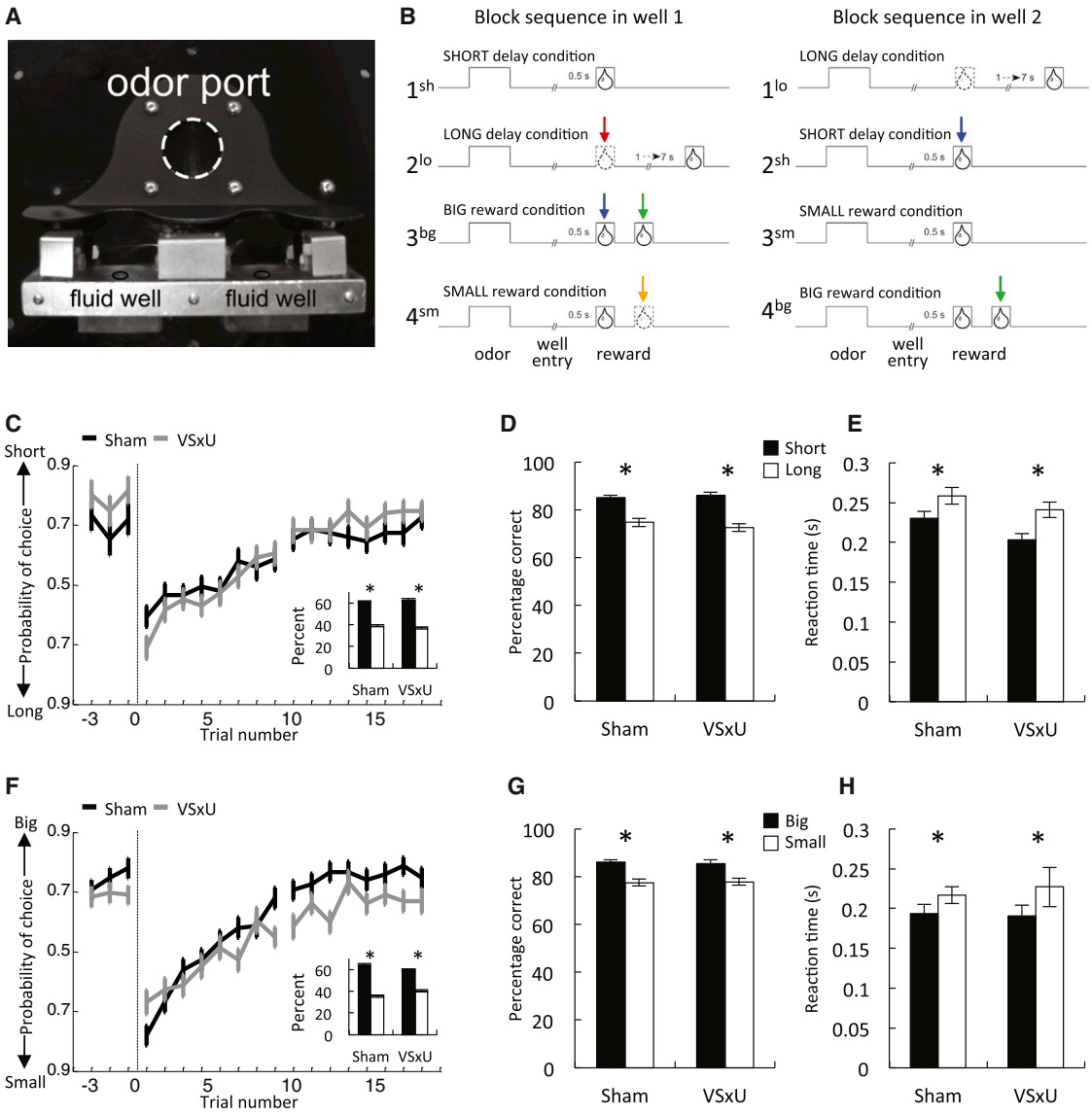

Figure 1. Apparatus and Behavioral Results

(A) Picture of apparatus used in the task, showing the odor port (~2.5 cm diameter) and two fluid wells.

(B) Line deflections indicate the time course of stimuli (odors and reward) presented to the animal on each trial. Dashed lines show when reward was omitted, and solid lines show when reward was delivered. At the start of each recording session, one well was randomly designated as short (a 0.5 s delay before reward) and the other, long (a 1–7 s delay before reward) (block 1). In the second block of trials, these contingencies were switched. In block 3, the delay was held constant while the number of reward was manipulated; one well was designated as big reward in which a second bolus of reward was delivered (big reward), and a single bolus of reward was delivered in the other well (small reward). In block 4, these contingencies were switched again. Blue arrows, unexpected short reward; red arrow, short reward omission; green arrows, unexpected big reward; orange arrow, big reward omission.

(C and F) Choice behavior in last three trials before the switch and first eight and last eight trials after the switch of reward contingencies in delay (C) and number blocks (F). Inset bar graphs show average percentage choice for high-valued (black) and low-valued (white) outcomes across all free-choice trials. Black line, sham-lesioned rats (Sham, n = 9, 75 sessions); gray line, unilateral VS-lesioned rats (VSxU, n = 7, 71 sessions).

(D, E, G, and H) Behavior on forced-choice trials in delay (D and E) and number blocks (G and H). Bar graphs show percentage correct (D and G) and reaction times (E and H) in response to the high and low value across all recording sessions. *p < 0.05 or better (see main text); NS, nonsignificant. Error bars represent SEM.

second odor signaled sucrose reward only in the right well (forced right), and a third odor signaled that reward was available at either well (free choice). To induce errors in the prediction of reward, we manipulated either the timing or the number of reward delivered in each well across blocks of trials (Figure 1B). Positive prediction errors were induced by making a previously delayed reward immediate (Figure 1B, blue arrows in blocks 2^{sh} and 3^{bg}) or by adding more reward (Figure 1B, green arrows in blocks 3^{bg} and 4^{bg}), whereas negative prediction errors were induced by delaying a previously immediate reward (Figure 1B,

red arrow in block 2^{lo}) or by decreasing the number of reward (Figure 1B, orange arrows in block 4^{sm}).

Recording began after the rats were shaped to perform the task. Shaping was similar across the two groups, and there were no significant differences in the number of trials in each block in the recording sessions (ANOVA, $F_{3,432} = 0.54$, $p = 0.66$). As expected, sham-lesioned rats changed their behavior across blocks in response to the changing reward, choosing the larger/earlier reward more often on free-choice trials (75 sessions; delay blocks, t test, $t_{74} = 7.96$, $p < 0.01$, Figure 1C; number blocks, t test, $t_{74} = 11.72$, $p < 0.01$, Figure 1F) and responding more accurately (delay blocks, t test, $t_{74} = 12.01$, $p < 0.01$, Figure 1D; number blocks, t test, $t_{74} = 9.29$, $p < 0.01$, Figure 1G) and with shorter reaction times (delay blocks, t test, $t_{74} = 5.81$, $p < 0.01$, Figure 1E; number blocks, t test, $t_{74} = 3.06$, $p < 0.01$, Figure 1H) on forced-choice trials when the earlier or larger reward was at stake. Rats with unilateral (VSxU) lesions showed similar behavior (71 sessions; percent choice in delay blocks, t test, $t_{70} = 12.81$, $p < 0.01$, Figure 1C; in number blocks, t test, $t_{70} = 8.29$, $p < 0.01$, Figure 1F; percent correct in delay blocks, t test, $t_{70} = 10.39$, $p < 0.01$, Figure 1D; in number blocks, t test, $t_{70} = 5.74$, $p < 0.01$, Figure 1G; reaction times in delay blocks, t test, $t_{70} = 7.03$, $p < 0.01$, Figure 1E; in number blocks, t test, $t_{70} = 3.06$, $p < 0.05$, Figure 1H). Two-factor ANOVAs (group × reward number or group × reward timing) revealed neither main effects nor any interactions involving group in free-choice performance, percent correct, or reaction times in delay ($F < 3.1$, $p > 0.08$) or number ($F < 0.7$, $p > 0.07$) blocks. Thus, the two groups showed similar differences in all our behavioral measures in both block types.

Dopamine Neurons Signal Prediction Errors in Response to Changes in Timing or Number of Reward

We identified putative dopamine neurons by means of a waveform analysis similar to that used to identify dopamine neurons in primate studies (Bromberg-Martin et al., 2010; Fiorillo et al., 2008; Hollerman and Schultz, 1998; Kobayashi and Schultz, 2008; Matsumoto and Hikosaka, 2009; Mirenowicz and Schultz, 1994; Morris et al., 2006; Waelti et al., 2001). Although the use of waveform criteria has not been uniformly accepted for isolating dopamine neurons (Margolis et al., 2006), this analysis isolates neurons in rat VTA whose firing is sensitive to intravenous infusion of apomorphine or quinpirole (Jo et al., 2013; Roesch et al., 2007), and nigral neurons identified by a similar cluster analysis in mice are selectively activated by optical stimulation in tyrosine hydroxylase channelrhodopsin-2 mutants and show reduced bursting in tyrosine hydroxylase striatal-specific NMDAR1 knockouts (Jin and Costa, 2010). Although these criteria may exclude some neurons containing enzymatic markers relevant to dopamine synthesis (Margolis et al., 2006; Ungless and Grace, 2012), only neurons in this cluster signaled reward prediction errors in appreciable numbers in our previous work (Roesch et al., 2007; Takahashi et al., 2011).

This approach identified as putatively dopaminergic 51 of 501 and 55 of 407 neurons recorded from VTA in sham- and VS-lesioned rats, respectively (Figures S1A–S1C). These proportions did not differ between groups (Chi-square = 2.4, df = 1, $p = 0.12$). Of these, 30 neurons in sham- and 31 in VS-lesioned rats increased firing in response to reward (compared with a 500-ms baseline taken during the inter-trial interval before trial onset). The average baseline activity and the average firing at the time of reward were similar in the two groups, both for these reward-responsive neurons as well as for the remaining dopamine neurons that were not responsive to reward (Figure S1D). Thus, VS lesions did not appear to have dramatic effects on the prevalence, waveform characteristics, or reward-related firing of putative dopamine neurons. Of note, neurons categorized as non-dopaminergic did show significantly higher baseline firing in the VS-lesioned rats (Figure S1E).

As in prior studies (Roesch et al., 2007; Takahashi et al., 2011), we found that prediction error signaling was largely restricted to reward-responsive, wide-waveform putative dopamine neurons (see Figure S2 for analyses of error signaling in other populations). In sham-lesioned rats, the activity of these neurons increased in response to an unexpected reward and decreased in response to omission of an expected reward. In each case, the change was largest at the start of the block, diminishing with learning of the new reward contingencies as the block proceeded. Firing to unexpected reward and its change with learning were similar whether we changed the timing (Figures 2A and 2E) or number of reward (Figures 2B and 2F). To quantify these effects, we analyzed the average firing rate on the first and last 10 trials in blocks in which we changed in the timing (Figure 2I) or number (Figure 2J) of reward. Activity was taken at the time of reward or reward omission in the relevant blocks, as indicated by the matching colored arrows in Figure 1B. A three-factor ANOVA (reward/omission × timing/number manipulation × trial) of the trial-by-trial neural data plotted in Figures 2I and 2J revealed main effects of reward/omission ($F_{1,29} = 15.0$, $p < 0.001$) and a significant interaction between reward/omission and trial ($F_{19,551} = 6.15$, $p < 0.001$), but no main effect nor any interactions involving timing/number manipulation ($F < 2.9$, $p > 0.10$). Separate ANOVAs indicated main effects of trial in each data series ($p < 0.01$), but no main effects or interactions with manipulation ($F < 2.7$, $p > 0.1$). Comparisons to baseline (gray lines, Figures 2I and 2J) revealed changes in firing initially in response to unexpected reward or reward omission in delay blocks (Figure 2I: reward/baseline × early/late phase, $F_{1,29} = 9.42$, $p < 0.01$; omission/baseline × early/late phase, $F_{1,29} = 9.59$, $p < 0.01$) and number blocks (Figure 2J: reward/baseline × early/late phase, $F_{1,29} = 8.97$, $p < 0.01$; omission/baseline × early/late phase, $F_{1,29} = 15.3$, $p < 0.01$). Post hoc analyses showed significant differences in firing versus baseline for both reward and omission early in each type of block (Figure 2I: reward, $F_{1,29} = 12.65$, $p < 0.01$: omission, $F_{1,29} = 18.9$, $p < 0.01$; Figure 2J: reward, $F_{1,29} = 8.11$, $p < 0.01$; omission, $F_{1,29} = 6.92$, $p < 0.05$), but not late in the block ($F < 3.8$, $p > 0.05$). In addition, difference scores comparing each neuron's firing early versus late in these blocks were distributed significantly above zero for unexpected reward (upper histograms in Figures 2I and 2J) and below zero for reward omission (lower histograms in Figures 2I and 2J).

Notably, the effects of changing the timing versus number of reward on the firing of these neurons were statistically indistinguishable inasmuch as the difference scores for unexpected reward and omission for each manipulation did not differ

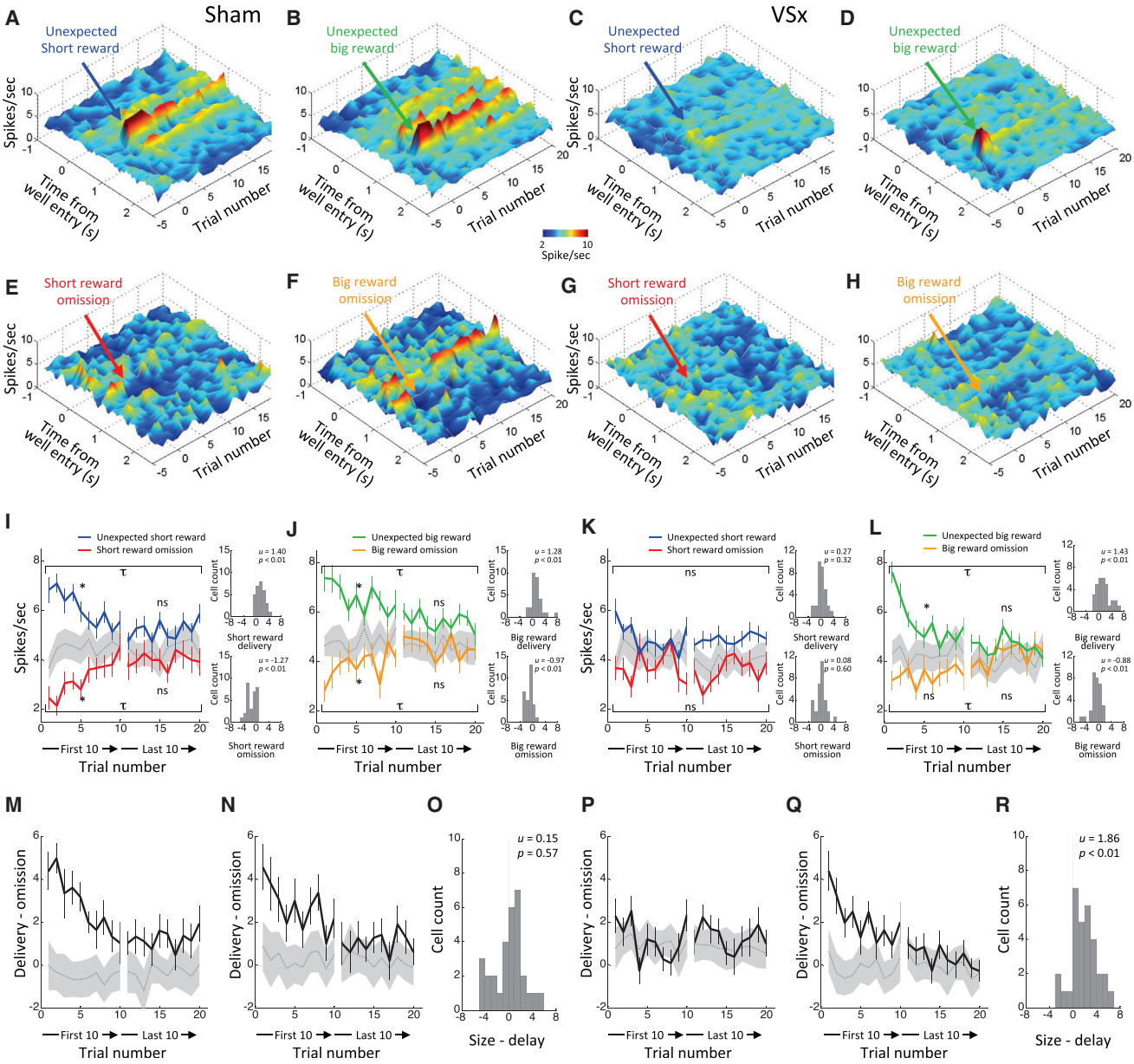

Figure 2. Changes in Activity of Reward-Responsive Dopamine Neurons to Unexpected Changes in Timing and Number of Reward

(A–H) Three-dimensional heat plots represent activity in averaged across all reward-responsive dopamine neurons in sham (n = 30) (A, B, E, and F) and VS-lesioned rats (n = 31) (C, D, G, and H) in response to introduction of unexpected delivery of short reward (A and C, blue arrows), unexpected big reward (B and D, green arrows), omission of expected short reward (E and G, red arrows), and omission of expected big reward (F and H, orange arrows).

(I–L) Average firing during 500 ms after delivery of short reward (blue) and big reward (green), or omission of short reward (red) and big reward (orange) in sham (I and J) and VS-lesioned rats (K and L). Error bars represent SEM. Gray dotted lines and gray shadings indicate baseline firing and SEM. t, significant interaction versus baseline; *, significant difference from baseline early or late; ns, non-significant. Small insets in each panel represent distribution of difference scores comparing firing to unexpected reward (top) and reward omission (bottom) in the early 5 versus late 10 trials in relevant trial blocks. Difference scores were computed from the average firing rate of each neuron in the first 5 minus the last 10 trials in relevant trial blocks. The numbers in the upper right of each panel indicate results of Wilcoxon signed-rank test (p) and the average difference score (u).

(M, N, P, and Q) Difference in firing between delivery and omission of short reward (M and P) and between delivery and omission of big reward (N and Q) in sham (M and N) and VS-lesioned rats (P and Q). Dashed lines and gray shadings indicate average and 2 SEM of shuffled data. Error bars represent SEM.

(O and R) Distribution of difference scores comparing (M) versus (N) in sham (O) and (P) versus (Q) in VS-lesioned rats (R). The numbers in upper right of each panel indicate results of Wilcoxon signed-rank test (p) and the average difference score (u).

(Wilcoxon rank-sum test, $p > 0.5$). The similarity was also apparent when we plotted the difference in firing upon delivery of unexpected reward versus reward omission, trial by trial, separately for changes in timing (Figure 2M) or number (Figure 2N). This difference was large initially in each data series and then diminished after a small number of trials, approaching a shuffled baseline. A two-factor ANOVA comparing these data across the two manipulations found a significant main effect of trial ($F_{19,551} = 6.23$, $p < 0.001$), but no main effect nor any interaction with manipulation ($F < 0.7$, $p > 0.75$, see also difference scores plotted in the histogram in Figure 2O). Together, these results suggest that our reward timing and number manipulations were equally successful at generating reward prediction errors in sham-lesioned animals.

VS Lesions Disrupt Dopamine Neuron Signaling of Prediction Errors in Response to Changes in Timing but Not Number of Reward

Ipsilateral lesions of the ventral striatum had a marked effect on the firing of dopamine neurons. In particular, reward-responsive dopamine neurons recorded in rats with ipsilateral VS lesions showed very little prediction error signaling when the timing of a reward changed (a delayed reward was made immediate, Figure 2C, or an immediate reward was delayed, Figure 2G). These neurons nevertheless showed prediction error signaling when reward number changed (new reward added, Figure 2D, or removed, Figure 2H), although the changes in firing also seemed somewhat muted compared to those observed in sham controls.

These effects were again quantified by analyzing the average firing of the reward-responsive dopamine neurons on the first and last 10 trials in all blocks in which we changed the timing (Figure 2K) or number (Figure 2L) of reward. A three-factor ANOVA (reward/omission × timing/number manipulation × trial number) of the neural data plotted in Figures 2K and 2L revealed main effects of reward/omission ($F_{1,30} = 15.4$, $p < 0.001$), trial ($F_{19,570} = 1.83$, $p < 0.05$), and a significant interaction between reward/omission and trial ($F_{19,570} = 3.46$, $p < 0.001$). However, in addition to these effects, which were similar to those seen in sham controls, there was also a significant three-way interaction involving the timing/number manipulation ($F_{19,570} = 2.66$, $p < 0.001$). Separate ANOVAs revealed significant interactions between trial and manipulation for both unexpected reward ($F_{19,570} = 1.91$, $p < 0.05$) and reward omission ($F_{19,570} = 1.67$, $p < 0.05$), and there were significant differences in firing on early versus late trials in response to changes in reward number ($p < 0.001$) but not reward timing ($F < 1.45$, $p > 0.10$). Accordingly, difference scores comparing each neuron's firing early versus late in blocks where reward number was changed were distributed significantly above zero for unexpected reward and below zero for reward omission (histograms, Figure 2L); however, similar scores computed when there was a change in reward timing were not different from zero (histograms, Figure 2K). The distributions of these scores also differed significantly across manipulations (Wilcoxon rank-sum tests, $p < 0.05$), indicating that the effects of changing the timing versus number of reward were statistically different.

Comparisons to baseline (gray lines, Figures 2K and 2L) generally confirmed this difference between the effects of the two manipulations. There were no changes in firing versus baseline in response to manipulation of reward timing (Figure 2K: reward/baseline × early/late phase, $F_{1,30} = 1.65$, $p = 0.21$; omission/baseline × early/late phase, $F_{1,30} = 0.79$, $p = 0.38$), whereas there were changes in response to manipulation of reward number (Figure 2L: reward/baseline × early/late phase, $F_{1,30} = 5.83$, $p < 0.05$; omission/baseline × early/late phase, $F_{1,30} = 18.9$, $p < 0.01$). Post hoc analyses showed significant differences in firing versus baseline early only for an unexpected increase in reward number (Figure 2K: reward, $F_{1,30} = 7.66$, $p < 0.01$).

The different effects of the two manipulations were particularly apparent when we plotted the difference in firing upon delivery of unexpected reward versus reward omission, trial by trial, separately for changes in timing (Figure 2P) or number (Figure 2Q) of reward. For changes in reward number, this difference was large initially, diminishing after a small number of trials to approach a shuffled baseline value. However, for changes in reward timing, this difference score was flat throughout the block, showing only a difference due to reward receipt or omission. A two-factor ANOVA comparing these data across the two manipulations found a significant interaction between trial and manipulation ($F_{19,570} = 2.65$, $p < 0.001$, see also difference scores plotted in the histogram in Figure 2R).

Finally, we directly compared data from sham- and VS-lesioned rats. ANOVAs comparing data for number blocks (Figure 2J versus Figure 2L or Figure 2N versus Figure 2Q) found no main effects nor any interactions involving group ($F < 1.3$, $p > 0.18$), with the distributions of the scores comparing firing early versus late in these blocks (histograms, Figure 2J versus Figure 2L) not statistically different across groups (Wilcoxon rank-sum test, addition: $p = 0.96$, omission: $p = 0.52$). These analyses indicate that, within the limits of our statistical power, the two groups showed very similar changes in firing in response to the addition or omission of extra reward. On the other hand, we found significant group interactions when comparing data from timing blocks in Figure 2I versus Figure 2K (reward/omission × trial × group: $F_{19,1121} = 2.63$, $p < 0.001$) and in Figure 2M versus Figure 2P (trial × group: $F_{19,1121} = 2.63$, $p < 0.001$). Post hoc analyses showed a significant interaction between group and firing on early versus late trials in response to both unexpected reward and reward omission ($p < 0.01$). Accordingly, the distributions of the difference scores comparing firing changes in these blocks (histograms, Figure 2I versus Figure 2K) were significantly different between the groups (Wilcoxon rank-sum tests, $p < 0.01$), reflecting the fact that dopamine neurons recorded in sham controls changed firing in response to changes in reward timing, whereas those recorded in VS-lesioned rats did not. Distributions of difference scores plotted in Figures 2O and 3R were also significantly different (Wilcoxon rank-sum test, $p = 0.011$). Together, these analyses show that putative dopamine neurons in VS-lesioned rats responded differently to changes in the timing of the reward compared to putative dopamine neurons recorded in sham-lesioned rats.

VS Provides Temporally Precise Reward Predictions to the VTA

The in vivo results of our experiment suggest that temporal aspects of reward prediction can be dissociated from predictions

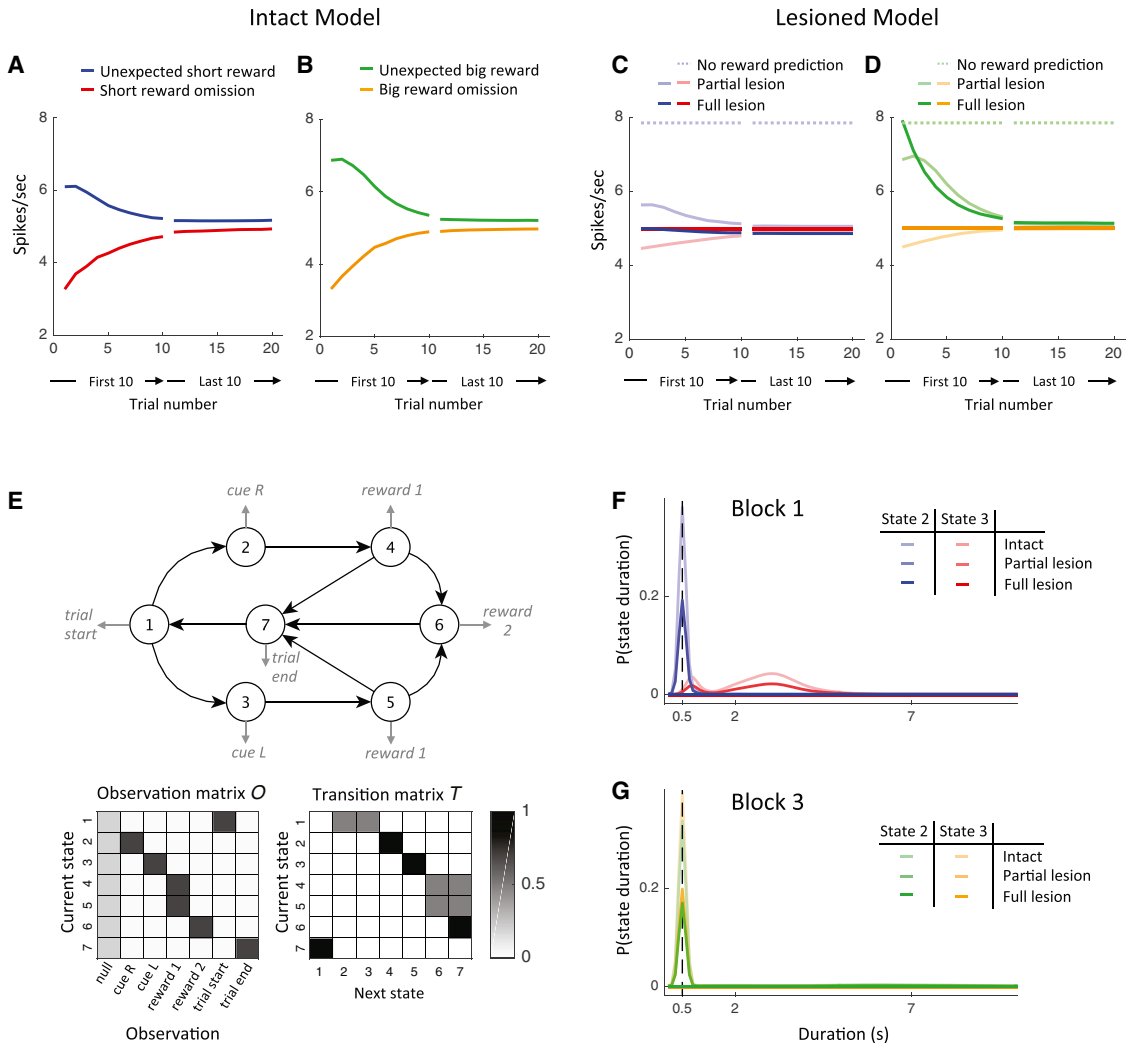

Figure 3. Effects of Simulating a Lesion of Temporal Expectations in the Semi-Markov TDRL Model

(A–D) Simulated average prediction errors during 500 ms after delivery of short reward (blue) and big reward (green), or omission of short reward (red) and big reward (orange) in the intact model (A and B) and in the lesioned model (C and D). Dotted lines in (C) and (D) show the expected pattern of prediction-error signals if total reward predictions were lost. Dark and light lines in (C) and (D) show simulated prediction-error signals for the full lesion and partial lesion models, respectively (colors indicate the same conditions as in A and B).

(E) State space representation of the task, with transitions between states marked in black arrows and the characteristic observation for each state marked in gray. Note that the characteristic observation is only emitted with p = 0.7; otherwise, the state emits a null (empty) observation (p = 0.2) or any of the other five possible observations (with p = 0.02 each). Below, the observation matrix shows the probability of each observation given each state and the transition matrix shows the probability of each successor state given each state.

(F) Learnt dwell-time distributions at the end of block 1 (delay block) for state 2 in the short delay condition (blue) and state 3 in the long delay condition (red) for the intact, partial lesion, and full lesion models.

(G) Similar to (F) but at the end of block 3 (size block) for state 2 in the big reward condition (green) and state 3 in the small reward condition (orange). The more severe the lesion, more of the probability mass lies at infinity (equivalent to no prediction of state duration).

of the number of expected reward and that VS supports essential information about the former, but perhaps not the latter, to VTA dopamine neurons. To make concrete this interpretation, we developed a TDRL model of the task using a framework that explicitly separates learning of reward number from learning to predict the timing of future reward. Briefly, rather than using the standard TDRL framework in which event identity and event timing are inextricable, we used a semi-Markov framework that represents and learns about reward timing (and, more generally, the duration of delays between events) separately and in parallel to representing and learning about expected amounts of reward (Daw et al., 2006).

In the model, the task is represented as a sequence of states (Figure 3E), with the duration of each state being potentially different. Key events in the task (e.g., appearance of an odor or a reward) signal transitions between states; however, transitions

can also occur without an external signal, due to the passage of time. Importantly, through experience with the task, the animal learns the expected duration for each state, as well as the value (i.e., expected amount of future reward) associated with the state. The latter is acquired through a temporal-difference learning rule, as in classic TDRL (see Experimental Procedures for full details), but it is separate from learning about state durations, which occurs through averaging of previous durations of each state. The separation of state durations from state values allows the model to capture disruption to one function but not the other (as evidenced in the dopamine signals) and consists of a departure from commonly held assumptions regarding the computational basis of prediction learning in the basal ganglia. That is, our model suggests a completely different implementation of reinforcement learning in basal ganglia circuitry than is commonly considered (Joel et al., 2002).

One important departure from classic TDRL is that in our model, prediction error signals occur only on state transitions and thus can be seen as "gated" according to the probability of a state transition occurring at any given time. This probability is 1 (or near 1, thus the "gate" is fully "open") when an external event is observed, as when a new reward is delivered. However, critically, the transition probability can also be high if there is a precise expectation regarding the duration of a specific state and this duration has terminated. It is in this case—when a state is deemed to terminate due to the passage of time despite the expected reward having not arrived—that a negative prediction error signal will be gated. Thus, temporal expectations can control or gate prediction errors, by causing a transition between states to be inferred even in the absence of an external event.

To test the predictions of the model, we simulated the evolution of state values, state durations, and the associated prediction errors using the actual task event timings that each group of rats experienced when performing the task. Model prediction-error signals were transformed into an instantaneous firing rate and then averaged and analyzed using the same epochs as we used to analyze the neural data. For the control group (full intact model), the simulations yielded the characteristic pattern of prediction error signals observed in vivo in this study and previously (Roesch et al., 2007; Takahashi et al., 2011). Specifically, the simulation produced positive prediction errors to unexpected early or large reward and negative prediction errors to unexpected reward omission (Figures 3A and 3B). These errors were strongest early in each block, gradually disappeared over subsequent trials, and the pattern was similar for errors induced by changes in number versus timing of reward.

To simulate a VS lesion in the model, we prevented the model from learning or using precise temporal expectations for the duration of states. To do this, updates of state durations were effectively "blurred" by decreasing the amplitude of the timing "kernel" that was used during learning (see Experimental Procedures and Figures 3F and 3G; partial lesions were modeled by decreasing the kernel by half, full lesions by decreasing it to baseline). This simulated lesion captures the intuition that the VS is critically involved in gating prediction errors according to learned state durations, and therefore a loss of VS corresponds specifically to (partially or fully) decreased amplitude of a signal that tracks the expectation of a reward-related event occurring at that point within a trial. One effect of this "lesion" is to effectively block the model from inferring a transition between states without an observation—the model will wait in a state indefinitely (or until the next observation of cue or reward) and is unable to infer a transition in the case of a missing (or late) reward.

This timing-lesioned model produced results that were remarkably similar to the firing of the dopamine neurons in VS-lesioned rats. Specifically, the simulation produced positive prediction errors in response to the delivery of new reward (Figure 3D, green line) but showed neither positive nor negative errors in response to changes in reward timing (Figure 3C). These results are identical to the data observed in vivo. Moreover, the model did not register negative prediction errors when the additional rewards were omitted in number blocks (Figure 3D, orange line). This is because the lesioned model could not use state duration to infer a transition at the time of the expected (but omitted) reward, and thus it did not gate a negative prediction-error signal. Notably, our neural data were equivocal on whether a negative prediction error occurred for this event. On the one hand, there was not a significant difference in firing to reward omission between groups and there appeared to be a significant shift below zero in the activity of the individual neurons at the time of omission in the number blocks. However, comparing firing of the dopamine neurons recorded in the lesioned rats at the time of reward omission to baseline at the start of these blocks, the apparent decline in firing was not statistically significant. In any event, any discrepancy here is not necessarily at odds with this prediction of the model, as it is possible that the lesions were not equivalent to a complete loss of function (see light lines in Figures 3C and 3D for simulation of the effects of a partial lesion). We also simulated a lesion in which the width of the kernel update to the state duration distributions was increased (rather than its amplitude decreased) and the calculation of expected duration within a state was left intact. This simulation produced similar results (Figure S3), suggesting that the specific implementation of the lesion was not paramount, as long as timing information in the model was degraded.

Finally we also tested the original hypothesis, commonly held in the literature (Joel et al., 2002; O'Doherty et al., 2003; Willuhn et al., 2012), that the VS serves as a (unitary) source of predictions of both when and how much reward is predicted for VTA dopamine neurons. To do this, in the lesioned model we set all state values to zero and blocked their update during the task, creating a situation where prediction errors must be computed absent input regarding both the timing and number of predicted reward. In this case, and as expected, the model generated persistent prediction errors to reward delivery and no prediction errors to reward omission (dotted lines in Figures 3C and 3D). This pattern of results is clearly at odds with the in vivo data, suggesting that the assumption embedded in classical TDRL models—that predictions of reward number and of reward timing go hand in hand (and thus are either present or absent as a unit)—is incorrect.

DISCUSSION

Reward prediction errors are signaled by midbrain dopamine neurons (Barto, 1995; Mirenowicz and Schultz, 1994; Montague

et al., 1996; Schultz et al., 1997). To do this, dopamine neurons require predictions to compare to actual obtained reward (Bush and Mosteller, 1951; Rescorla and Wagner, 1972; Sutton and Barto, 1998). Theoretical and experimental work has suggested that the VS is an important source of these predictions, particularly to dopamine neurons in the VTA (Daw et al., 2005, 2006; Joel et al., 2002; O'Doherty et al., 2003, 2004; Seymour et al., 2004). Here we tested this hypothesis, recording from VTA dopamine neurons in rats with a lesioned VS, while they performed a task in which positive and negative prediction errors were induced by shifting either the timing or the number of expected reward. Sham-lesioned rats exhibited prediction error signals in response to both manipulations, replicating our previous findings (Roesch et al., 2007; Takahashi et al., 2011). By contrast, dopamine neurons in rats with ipsilateral VS lesions exhibited intact prediction errors in response to increases in the number of reward but no prediction errors to changes in reward timing on the order of several seconds (and possibly also to complete omission of expected reward). These effects were reproduced by a computational model that used a non-traditional reinforcement-learning framework to separate learning about reward timing from learning about reward number ("state value"). Our results thus suggest a critical role for the VS in providing information about the predicted timing of reward, but not their number, to VTA dopamine neurons. These data and our theoretical interpretive framework may require a rethinking of the implementation of prediction-error computation in the neural circuitry of the basal ganglia.

Before considering the implications of our findings, we address some important caveats. One key determinant of our findings may be the amount of training the rats underwent before recording began—while our rats were trained extensively, it may be that the VS has a broader role in supporting reward predictions in very early stages of learning a task. It is also possible that had we allowed less time for compensation by using reversible inactivation, or lesioned the VS more completely or bilaterally, we might have observed effects of the lesion on both types of prediction errors. In any case, our data suggest that delay-induced prediction-error signaling is more sensitive to VS damage than are prediction errors induced by changes in number of reward. We also failed to observe any relationship between the amount of damage and the loss or preservation of prediction errors in our lesioned group (Figure 2R versus % damage yields $r = 0.08$, $p = 0.66$). Even relatively modest damage to VS was sufficient to entirely disrupt delay-induced errors with little effect on those induced by number changes.

It also an empirical question of whether these results will generalize to primates or other rodent species or to other midbrain regions. While the waveform sorting approach we use is roughly similar to the approach used to identify prediction-error signaling dopamine neurons in primates (Bromberg-Martin et al., 2010; Fiorillo et al., 2008; Hollerman and Schultz, 1998; Kobayashi and Schultz, 2008; Matsumoto and Hikosaka, 2009; Mirenowicz and Schultz, 1994; Morris et al., 2006; Waelti et al., 2001), and the error signals we find in our neurons obey many of the same rules as those demonstrated in comparable primate studies (Bromberg-Martin et al., 2010; Fiorillo et al., 2008; Kobayashi and Schultz, 2008; Morris et al., 2006), it is possible that the influence of VS on these signals may be different in other species. This may be particularly true in mice, where dopamine neurons seem more prevalent in single-unit data (>50% of recorded neurons) and prediction-error signals are often reported in populations whose waveform features are largely indistinguishable from the other populations (Cohen et al., 2012; Eshel et al., 2015; Tian and Uchida, 2015). It is also possible that the influence of VS on dopaminergic error signals in other parts of midbrain differs from what we have observed in (mostly lateral) VTA.

As a last caveat, we note that our conclusions are based on a relatively small proportion of our recorded neurons, smaller than would be identified as dopaminergic by immunohistological criteria (Li et al., 2013). While we did not pre-select neurons for recording, it is possible that neurons with different firing correlates were not visible to our electrodes. Nevertheless, the neurons we isolated had waveform and firing correlates similar to putative dopamine neurons in other studies in both rats (Jo et al., 2013; Jo and Mizumori, 2015; Pan et al., 2005) and primates (Bromberg-Martin et al., 2010; Fiorillo et al., 2008; Kobayashi and Schultz, 2008; Morris et al., 2006), recorded in both VTA and substantia nigra pars compacta. Further, most neurons in lateral VTA of the rat are thought to be classical dopamine neurons, meaning that they exhibit the same enzymatic phenotype characteristic of dopamine-releasing neurons in the substantia nigra (Li et al., 2013). As a result, we do not believe we are missing substantial numbers of dopamine neurons in our classification. Consistent with this, we also analyzed activity from the other neural populations that we recorded, but did not see any evidence of significant dopamine-like error signaling (see Supplemental Information, especially Table S1).

We now turn to the possible implications of our results. Most importantly, our findings are inconsistent with the currently popular model in which VS supplies the reward predictions used by VTA dopamine neurons to calculate reward prediction errors (Daw et al., 2005, 2006; Joel et al., 2002; O'Doherty et al., 2003, 2004; Seymour et al., 2004). If this model were correct, removing VS input to dopamine neurons would have amounted to removing all reward predictions, and therefore would have resulted in persistent firing to all reward (now "unexpected") and no signals to omission of reward, irrespective of the manipulation (timing or number) that induced prediction errors. We did not observe these results, suggesting that the strong version of this hypothesis is not viable.

However, we did find major effects of VS removal on VTA error signaling, but only when prediction errors were induced by changing the timing of the reward. Here it is important to note that even for these timing-dependent prediction errors, this is not the effect we would have expected if we simply eliminated timing predictions—while negative prediction errors would be lost, positive errors would remain high in that case, as all reward would be surprising. Instead, we found that positive errors were also eliminated (as compared to the response of these neurons to other unexpected reward), as if a reward was predicted *whenever it arrived*. That is, lacking VS input, putative dopamine neurons knew that reward would appear but did not know (or care) when. This suggests that VS is not necessary for predicting the occurrence of reward; however it is necessary for endowing that prediction with temporal specificity (e.g., that the reward is

coming after 500 ms). Intriguingly, absent this information, *any* timing of the reward was treated as the expected timing of the reward.

A dissociation between knowing *that* reward is coming but not *when* it should arrive directly contradicts the standard TDRL framework in which predictions about reward timing and number are inextricably linked (Montague et al., 1996; Schultz et al., 1997; Sutton and Barto, 1998). However, this dissociation was produced effectively in a model based on a semi-Markov framework proposed by Daw et al. (2006) that learns about reward timing and reward number in parallel. ''Lesioning'' this model removing expectations for the duration between events in the task and leaving all other aspects of prediction learning intact produced the exact pattern of altered prediction-error signaling seen in the VS-lesioned rats: a reward that came too early or was delayed such that its timing became uncertain produced no prediction error, even though there were robust prediction error signals in response to changes in number of reward. The model also predicted the loss of number-induced negative prediction errors (Figure 3D, orange line). It is not clear from our in vivo data whether this signal is indeed affected by the VS lesion (Figure 2J versus Figure 2L, orange lines), as the low baseline rates of dopamine neurons make suppression of firing (or lack thereof) difficult to demonstrate reliably. In any case, if one assumes that the lesions did not completely ablate the VS, some residual negative prediction error might be expected (Figures 3C and 3D, light lines).

Our modeling results link signaling from the VS with the gating of prediction-error signals according to the learned timing of reward, suggesting that activity in the VS might evolve within a trial by tracking learned temporal expectations regarding the timing of reward delivery (or other reward-predictive events). Such a signal from the VS might thus be similar to adaptive time representations found in the dorsal striatum in an instrumental task (Mello et al., 2015). A role for VS in constraining reward predictions to specific (learned) time intervals is also consistent with previous reports that place the VS at the center of a neural timing circuit (Meck et al., 2008). Notably this function is thought to depend on input from hippocampus (Meck, 1988; Meck et al., 2013), which has been linked to keeping track of internal representations of time (Eichenbaum, 2014) and is known to regulate VS input to VTA (Floresco et al., 2001). The loss of the ability to track these temporal predictions absent a VS and its effect on the perception of negative prediction errors may also be of relevance to the apparent role of VS in the perception of risk (Dalton et al., 2014; St Onge et al., 2012; Stopper and Floresco, 2011), since an inability to perceive negative prediction errors would dramatically reduce the apparent ''riskiness'' of a probabilistic reward. One prediction of our model is that substituting a small reward for reward omission might restore normal function in such tasks.

A specific role of VS in signaling state timing can even be observed in simple Pavlovian conditioning tasks—previous work has shown that when rats with bilateral neurotoxic lesions of VS are trained to associate long-duration cues with reward, learning and the ultimate levels of responding are not affected by the lesions, but the specific pattern of responding *during* the cues is disrupted (Singh et al., 2011). Specifically, while non-lesioned controls exhibit progressively more conditioned responding as the reward period approached, rats with VS lesions respond at constant levels throughout the delay to reward. This is consistent with a failure to learn temporally precise predictions but a conserved ability to learn that a reward is impending. Finally, Klein-Flügge et al. (2011) compared fMRI responses in VS and VTA in humans learning to predict the timing, but not the number of variably timed reward. They showed that blood-oxygen-level-dependent (BOLD) signals in the VS are sensitive to information about timing, but not information about number, whereas VTA prediction-error signals are sensitive to both. Indeed, VS signals in that study were consistent with a role for the VS in learning when reward will occur, showing larger responses to cues that predicted information about timing and to reward that occurred within the expected timeframe. Here we have shown more directly that the VS supplies information about temporal specificity of predictions to VTA neurons.

Single-unit studies show that signals related to reward number or value are present in the VS, and many other reports indicate that VS integrity can be critical to behaviors seemingly based on value or even reward number (Berridge and Robinson, 1998; Di Chiara, 2002; Hauber et al., 2000; McDannald et al., 2011; Nicola, 2007; Steinberg et al., 2014). In fact, single-unit data from VS in the same task as used here show that information about number is present in VS unit activity (Roesch et al., 2009), and rats with bilateral VS lesions tested in the same task used in our study initially showed free-choice deficits (Burton et al., 2014), though the deficits in delay blocks were significantly more severe than those in the number blocks. Indeed, when analyzed separately, VS lesioned rats showed a preference for larger reward on free-choice trials but did not show a preference between immediate versus delayed reward (M.R. Roesch, personal communication), consistent with our finding. Interestingly, with further training, rats with bilateral VS lesions did learn to respond normally, even in the delay blocks. The authors concluded that this recovery of normal behavior must reflect compensation by other slower learning mechanisms in dorsal striatum (Burton et al., 2014). Our data from extensively trained rats suggest that these mechanisms can operate independent of normal delay-induced dopaminergic error signals from the VTA.

Our data suggest that even though VS neurons may signal information about the expected number of reward (Roesch et al., 2009), this information may be redundant with inputs to VTA from other areas, as VS lesions were not detrimental to (positive) prediction-error signaling due to changes in reward number. Indeed, information about reward number or simply the future occurrence of reward is sufficiently fundamental that it is likely signaled by a diverse array of areas impinging on the midbrain, any one of which may be sufficient to support error signaling in response to simply adding a reward. Neural responses to reward or to cues that predict different numbers of reward are found in many brain areas, including areas that have either direct or indirect projections to VTA or its surrounds. Preserved number-induced error signals in our recordings may therefore reflect input from any of these areas. By contrast, our results suggest that signaling of *when* a reward is expected to occur is an aspect of reward prediction that is mediated uniquely by circuitry that converges on the VS.

EXPERIMENTAL PROCEDURES

Subjects

Male Long-Evans rats (n = 16) were obtained at 175–200 g from Charles River Laboratories. Rats were tested at the NIDA-IRP in accordance with NIH guidelines.

Surgical Procedures and Histology

Lesions were made and electrodes implanted under stereotaxic guidance; all surgical procedures adhered to guidelines for aseptic technique. VS lesions were made by infusing of quinolinic acid (Sigma) in Dulbecco's phosphate vehicle. Infusions of 0.4 μl quinolinic acid (20 μg μl^{-1}) were made at 1.9 mm anterior to the bregma, and 1.9 mm lateral to the midline, at a depth of 7.3 mm ventral to the skull surface. Sham controls received identical treatment only no infusion was made. After this procedure, a drivable bundle of eight 25-um diameter FeNiCr wires (Stablohm 675, California Fine Wire) chronically implanted dorsal to VTA in the left or right hemisphere at 5.2 mm posterior to bregma, 0.7 mm laterally, and 7.0 mm ventral to the brain surface at an angle of 5° toward the midline from vertical. Wires were cut with surgical scissors to extend ~1.5 mm beyond the cannula and electroplated with platinum (H_2PtCl_6, Aldrich) to an impedance of ~300 kOhms. Cephalexin (15 mg/kg by mouth [p.o.]) was administered twice daily for 2 weeks post-operatively. The rats were then perfused, and their brains removed and processed for histology (Roesch et al., 2006).

Odor-Guided Choice Task

Recording was conducted in aluminum chambers approximately 18 in on each side with sloping walls narrowing to an area of 12 in × 12 in at the bottom. A central odor port was located above two fluid wells (Figure 1A). Two lights were located above the panel. The odor port was connected to an air flow dilution olfactometer to allow the rapid delivery of olfactory cues. Odors where chosen from compounds obtained from International Flavors and Fragrances.

Trials were signaled by illumination of the panel lights inside the box. When these lights were on, nosepoke into the odor port resulted in delivery of the odor cue to a small hemicylinder located behind this opening. One of three different odors was delivered to the port on each trial, in a pseudorandom order. At odor offset, the rat had 3 s to make a response at one of the two fluid wells. One odor instructed the rat to go to the left to get reward, a second odor instructed the rat to go to the right to get reward, and a third odor indicated that the rat could obtain reward at either well. Odors were presented in a pseudorandom sequence such that the free-choice odor was presented on 7/20 trials and the left/right odors were presented in equal numbers. In addition, the same odor could be presented on no more than 3 consecutive trials.

Once the rats were shaped to perform this basic task, we introduced blocks in which we independently manipulated the number of the reward and the delay preceding reward delivery (Figure 1B). For recording, one well was randomly designated as short and the other long at the start of the session (Figure 1B, 1^{sh} and 1^{lo}). In the second block of trials, these contingencies were switched (Figure 1B, 2^{sh} and 2^{lo}). The length of the delay under long conditions followed an algorithm in which the side designated as long started off as 1 s and increased by 1 s every time that side was chosen until it became 3 s. If the rat continued to choose that side, the length of the delay increased by 1 s up to a maximum of 7 s. If the rat chose the side designated as long less than 8 out of the last 10 choice trials, then the delay was reduced by 1 s to a minimum of 3 s. The reward delay for long forced-choice trials was yoked to the delay in free-choice trials during these blocks. In later blocks we held the delay preceding reward constant while manipulating the number of reward (Figure 1B, 3^{bg}, 3^{sm}, 4^{bg}, and 4^{sm}). The reward was a 0.05 ml bolus of 10% sucrose solution. The reward number used in delay blocks was the same as the reward used in the small reward blocks. For big reward, additional boli were delivered after gaps of 500 ms.

Single-Unit Recording

Wires were screened for activity daily; if no activity was detected, the rat was removed, and the electrode assembly was advanced 40 or 80 um. Otherwise active wires were selected to be recorded, a session was conducted, and the electrode was advanced at the end of the session. Neural activity was recorded using Plexon Multichannel Acquisition Processor systems. Signals from the electrode wires were amplified 20× by an op-amp headstage (Plexon, HST/8o50-G20-GR), located on the electrode array. Immediately outside the training chamber, the signals were passed through a differential preamplifier (Plexon, PBX2/16sp-r-G50/16fp-G50), where the single-unit signals were amplified 50× and filtered at 150–9,000 Hz. The single-unit signals were then sent to the Multichannel Acquisition Processor box, where they were further filtered at 250–8,000 Hz, digitized at 40 kHz, and amplified at 1–32×. Waveforms (>2.5:1 signal-to-noise) were extracted from active channels and recorded to disk by an associated workstation.

Data Analysis

Units were sorted using Offline Sorter software from Plexon. Sorted files were then processed and analyzed in Neuroexplorer and MATLAB. Dopamine neurons were identified via a waveform analysis used and validated by us and others previously (Jo et al., 2013; Roesch et al., 2007; Takahashi et al., 2011; Jin and Costa, 2010). Briefly, cluster analysis was performed based on the half time of the spike duration and the ratio comparing the amplitude of the first positive and negative waveform segments. The center and variance of each cluster was computed without data from the neuron of interest, and then that neuron was assigned to a cluster if it was within 3 SD of the cluster's center. Neurons that met this criterion for more than one cluster were not classified. This process was repeated for each neuron.

To quantify changes in firing due to reward delivery or omission, we examined activity in the 500-ms periods identified by the arrows in Figure 1B. The start of this time window coincided with opening of the solenoid valve, and the duration was chosen to encompass the maximum duration of opening (actual open times were calibrated to maintain 0.05 ml boli and so could be less than this). Importantly, no other trial event occurred until at least 500 ms after the end of this time window. Thus, this period allowed us to isolate activity related to delivery or omission of each individual reward bolus. Analyses were conducted using MATLAB (MathWorks) or Statistica (StatSoft) as described in the main text.

Computational Modeling

We simulated learning and prediction error signaling in the task using TDRL in a semi-Markov framework with partial observability (Daw et al., 2006). Briefly, in this approach, we assume the rats represent the behavioral task as a sequence of states that each have an associated value, *V*, and a distribution over dwell times in that state, *D*. Observations during the task, such as an odor cue or the delivery of a reward in a well, signal a transition between states, at which time a prediction error is signaled and used to update state values. Additionally, transitions can occur without an external observation due to the mere passage of time (e.g., at the time that a reward was expected but failed to arrive, see below). That is, knowledge of the likely dwell time in a state (represented by *D*) can be used to infer a silent transition and gate the signaling of a prediction error and the update of state values. To simulate a VS lesion in the model, we prevented the model from learning accurate dwell time distributions for each state, thereby degrading the ability of the model to infer these silent transitions when a reward is omitted or delayed. We describe the model in more detail in the Supplemental Information.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, three figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2016.05.015.

AUTHOR CONTRIBUTIONS

Y.K.T. and G.S. conceived and designed the behavioral and single-unit experiments, and Y.K.T. conducted the experiments and analyzed the data. A.J.L. conducted the modeling, with input from Y.N. Y.K.T. and G.S. prepared the manuscript with input from the other authors.

ACKNOWLEDGMENTS

This work was supported by funding from NIDA (to Y.K.T. and G.S.), the Human Frontier Science Program Organization (to A.J.L.), and NIMH grant R01MH098861 (to Y.N.). The opinions expressed in this article are the authors' own and do not reflect the view of the NIH, the Department of Health and Human Services, or the United States government. The authors would like to acknowledge Nathaniel Daw for helpful suggestions regarding the semi-Markov model of the task.

Received: December 22, 2015
Revised: March 30, 2016
Accepted: April 27, 2016
Published: June 9, 2016

REFERENCES

Barto, A.G. (1995). Adaptive critics and the basal ganglia. In Models of Information Processing in the Basal Ganglia, J.C. Houk, J.L. Davis, and D.G. Beiser, eds. (MIT Press), pp. 215–232.

Berridge, K.C., and Robinson, T.E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Res. Brain Res. Rev. *28*, 309–369.

Bissonette, G.B., Burton, A.C., Gentry, R.N., Goldstein, B.L., Hearn, T.N., Barnett, B.R., Kashtelyan, V., and Roesch, M.R. (2013). Separate populations of neurons in ventral striatum encode value and motivation. PLoS ONE *8*, e64673.

Bocklisch, C., Pascoli, V., Wong, J.C.Y., House, D.R.C., Yvon, C., de Roo, M., Tan, K.R., and Lüscher, C. (2013). Cocaine disinhibits dopamine neurons by potentiation of GABA transmission in the ventral tegmental area. Science *341*, 1521–1525.

Bromberg-Martin, E.S., Matsumoto, M., Hong, S., and Hikosaka, O. (2010). A pallidus-habenula-dopamine pathway signals inferred stimulus values. J. Neurophysiol. *104*, 1068–1076.

Burton, A.C., Bissonette, G.B., Lichtenberg, N.T., Kashtelyan, V., and Roesch, M.R. (2014). Ventral striatum lesions enhance stimulus and response encoding in dorsal striatum. Biol. Psychiatry *75*, 132–139.

Bush, R.R., and Mosteller, F. (1951). A mathematical model for simple learning. Psychol. Rev. *58*, 313–323.

Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., and Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature *482*, 85–88.

Dalton, G.L., Phillips, A.G., and Floresco, S.B. (2014). Preferential involvement by nucleus accumbens shell in mediating probabilistic learning and reversal shifts. J. Neurosci. *34*, 4618–4626.

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

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.

Di Chiara, G. (2002). Nucleus accumbens shell and core dopamine: differential role in behavior and addiction. Behav. Brain Res. *137*, 75–114.

Eichenbaum, H. (2014). Time cells in the hippocampus: a new dimension for mapping memories. Nat. Rev. Neurosci. *15*, 732–744.

Eshel, N., Bukwich, M., Rao, V., Hemmelder, V., Tian, J., and Uchida, N. (2015). Arithmetic and local circuitry underlying dopamine prediction errors. Nature *525*, 243–246.

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

Floresco, S.B., Todd, C.L., and Grace, A.A. (2001). Glutamatergic afferents from the hippocampus to the nucleus accumbens regulate activity of ventral tegmental area dopamine neurons. J. Neurosci. *21*, 4915–4922.

Grace, A.A., and Bunney, B.S. (1985). Opposing effects of striatonigral feedback pathways on midbrain dopamine cell activity. Brain Res. *333*, 271–284.

Groenewegen, H.J., Berendse, H.W., Wolters, J.G., and Lohman, A.H.M. (1990). The anatomical relationship of the prefrontal cortex with the striatopallidal system, the thalamus and the amygdala: evidence for a parallel organization. Prog. Brain Res. *85*, 95–116, discussion 116–118.

Hauber, W., Bohn, I., and Giertler, C. (2000). NMDA, but not dopamine D(2), receptors in the rat nucleus accumbens areinvolved in guidance of instrumental behavior by stimuli predicting reward magnitude. J. Neurosci. *20*, 6282–6288.

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.

Jin, X., and Costa, R.M. (2010). Start/stop signals emerge in nigrostriatal circuits during sequence learning. Nature *466*, 457–462.

Jo, Y.S., and Mizumori, S.J. (2015). Prefrontal regulation of neuronal activity in the ventral tegmental area. Cereb. Cortex, bhv215, epub ahead of print.

Jo, Y.S., Lee, J., and Mizumori, S.J. (2013). Effects of prefrontal cortical inactivation on neural activity in the ventral tegmental area. J. Neurosci. *33*, 8159–8171.

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.

Klein-Flügge, M.C., Hunt, L.T., Bach, D.R., Dolan, R.J., and Behrens, T.E.J. (2011). Dissociable reward and timing signals in human midbrain and ventral striatum. Neuron *72*, 654–664.

Kobayashi, S., and Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. J. Neurosci. *28*, 7837–7846.

Li, X., Qi, J., Yamaguchi, T., Wang, H.-L., and Morales, M. (2013). Heterogeneous composition of dopamine neurons of the rat A10 region: molecular evidence for diverse signaling properties. Brain Struct. Funct. *218*, 1159–1176.

Margolis, E.B., Lock, H., Hjelmstad, G.O., and Fields, H.L. (2006). The ventral tegmental area revisited: is there an electrophysiological marker for dopaminergic neurons? J. Physiol. *577*, 907–924.

Matsumoto, M., and Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature *459*, 837–841.

McDannald, M.A., Lucantonio, F., Burke, K.A., Niv, Y., and Schoenbaum, G. (2011). Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. J. Neurosci. *31*, 2700–2705.

Meck, W.H. (1988). Hippocampal function is required for feedback control of an internal clock's criterion. Behav. Neurosci. *102*, 54–60.

Meck, W.H., Penney, T.B., and Pouthas, V. (2008). Cortico-striatal representation of time in animals and humans. Curr. Opin. Neurobiol. *18*, 145–152.

Meck, W.H., Church, R.M., and Olton, D.S. (2013). Hippocampus, time, and memory. Behav. Neurosci. *127*, 642–654.

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

Mirenowicz, J., and Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. J. Neurophysiol. *72*, 1024–1027.

Mogenson, G.J., Jones, D.L., and Yim, C.Y. (1980). From motivation to action: functional interface between the limbic system and the motor system. Prog. Neurobiol. *14*, 69–97.

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.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. Nat. Neurosci. *9*, 1057–1063.

Nicola, S.M. (2007). The nucleus accumbens as part of a basal ganglia action selection circuit. Psychopharmacology (Berl.) *191*, 521–550.

O'Doherty, J., Dayan, P., Friston, K.J., Critchley, H., and Dolan, R.J. (2003). Temporal difference models and reward-related learning in the human brain. Neuron *38*, 329–337.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R.J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science *304*, 452–454.

Oleson, E.B., Beckert, M.V., Morra, J.T., Lansink, C.S., Cachope, R., Abdullah, R.A., Loriaux, A.L., Schetters, D., Pattij, T., Roitman, M.F., et al. (2012). Endocannabinoids shape accumbal encoding of cue-motivated behavior via CB1 receptor activation in the ventral tegmentum. Neuron *73*, 360–373.

Pan, W.-X., Schmidt, R., Wickens, J.R., and Hyland, B.I. (2005). Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J. Neurosci. *25*, 6235–6242.

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, A.H. Black and W.F. Prokasy, eds. (Appleton-Century-Crofts), pp. 64–99.

Roesch, M.R., Taylor, A.R., and Schoenbaum, G. (2006). Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation. Neuron *51*, 509–520.

Roesch, M.R., Calu, D.J., and Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat. Neurosci. *10*, 1615–1624.

Roesch, M.R., Singh, T., Brown, P.L., Mullins, S.E., and Schoenbaum, G. (2009). Ventral striatal neurons encode the value of the chosen action in rats deciding between differently delayed or sized rewards. J. Neurosci. *29*, 13365–13376.

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

Seymour, B., O'Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J., and Frackowiak, R.S. (2004). Temporal difference models describe higher-order learning in humans. Nature *429*, 664–667.

Singh, T., McDannald, M.A., Takahashi, Y.K., Haney, R.Z., Cooch, N.K., Lucantonio, F., and Schoenbaum, G. (2011). The role of the nucleus accumbens in knowing when to respond. Learn. Mem. *18*, 85–87.

St Onge, J.R., Stopper, C.M., Zahm, D.S., and Floresco, S.B. (2012). Separate prefrontal-subcortical circuits mediate different components of risk-based decision making. J. Neurosci. *32*, 2886–2899.

Steinberg, E.E., Boivin, J.R., Saunders, B.T., Witten, I.B., Deisseroth, K., and Janak, P.H. (2014). Positive reinforcement mediated by midbrain dopamine neurons requires D1 and D2 receptor activation in the nucleus accumbens. PLoS ONE *9*, e94771.

Stopper, C.M., and Floresco, S.B. (2011). Contributions of the nucleus accumbens and its subregions to different aspects of risk-based decision making. Cogn. Affect. Behav. Neurosci. *11*, 97–112.

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

Takahashi, Y.K., Roesch, M.R., Stalnaker, T.A., Haney, R.Z., Calu, D.J., Taylor, A.R., Burke, K.A., and Schoenbaum, G. (2009). The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. Neuron *62*, 269–280.

Takahashi, Y.K., Roesch, M.R., Wilson, R.C., Toreson, K., O'Donnell, P., Niv, Y., and Schoenbaum, G. (2011). Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. Nat. Neurosci. *14*, 1590–1597.

Tian, J., and Uchida, N. (2015). Habenula lesions reveal that multiple mechanisms underlie dopamine prediction errors. Neuron *87*, 1304–1316.

Ungless, M.A., and Grace, A.A. (2012). Are you or aren't you? Challenges associated with physiologically identifying dopamine neurons. Trends Neurosci. *35*, 422–430.

Voorn, P., Vanderschuren, L.J.M.J., Groenewegen, H.J., Robbins, T.W., and Pennartz, C.M.A. (2004). Putting a spin on the dorsal-ventral divide of the striatum. Trends Neurosci. *27*, 468–474.

Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. Nature *412*, 43–48.

Watabe-Uchida, M., Zhu, L., Ogawa, S.K., Vamanrao, A., and Uchida, N. (2012). Whole-brain mapping of direct inputs to midbrain dopamine neurons. Neuron *74*, 858–873.

Willuhn, I., Burgeno, L.M., Everitt, B.J., and Phillips, P.E.M. (2012). Hierarchical recruitment of phasic dopamine signaling in the striatum during the progression of cocaine use. Proc. Natl. Acad. Sci. USA *109*, 20703–20708.

Xia, Y., Driscoll, J.R., Wilbrecht, L., Margolis, E.B., Fields, H.L., and Hjelmstad, G.O. (2011). Nucleus accumbens medium spiny neurons target non-dopaminergic neurons in the ventral tegmental area. J. Neurosci. *31*, 7811–7816.

Neuron, Volume *91*

Supplemental Information

Temporal Specificity of Reward Prediction Errors Signaled by Putative Dopamine Neurons in Rat VTA Depends on Ventral Striatum

Yuji K. Takahashi, Angela J. Langdon, Yael Niv, and Geoffrey Schoenbaum

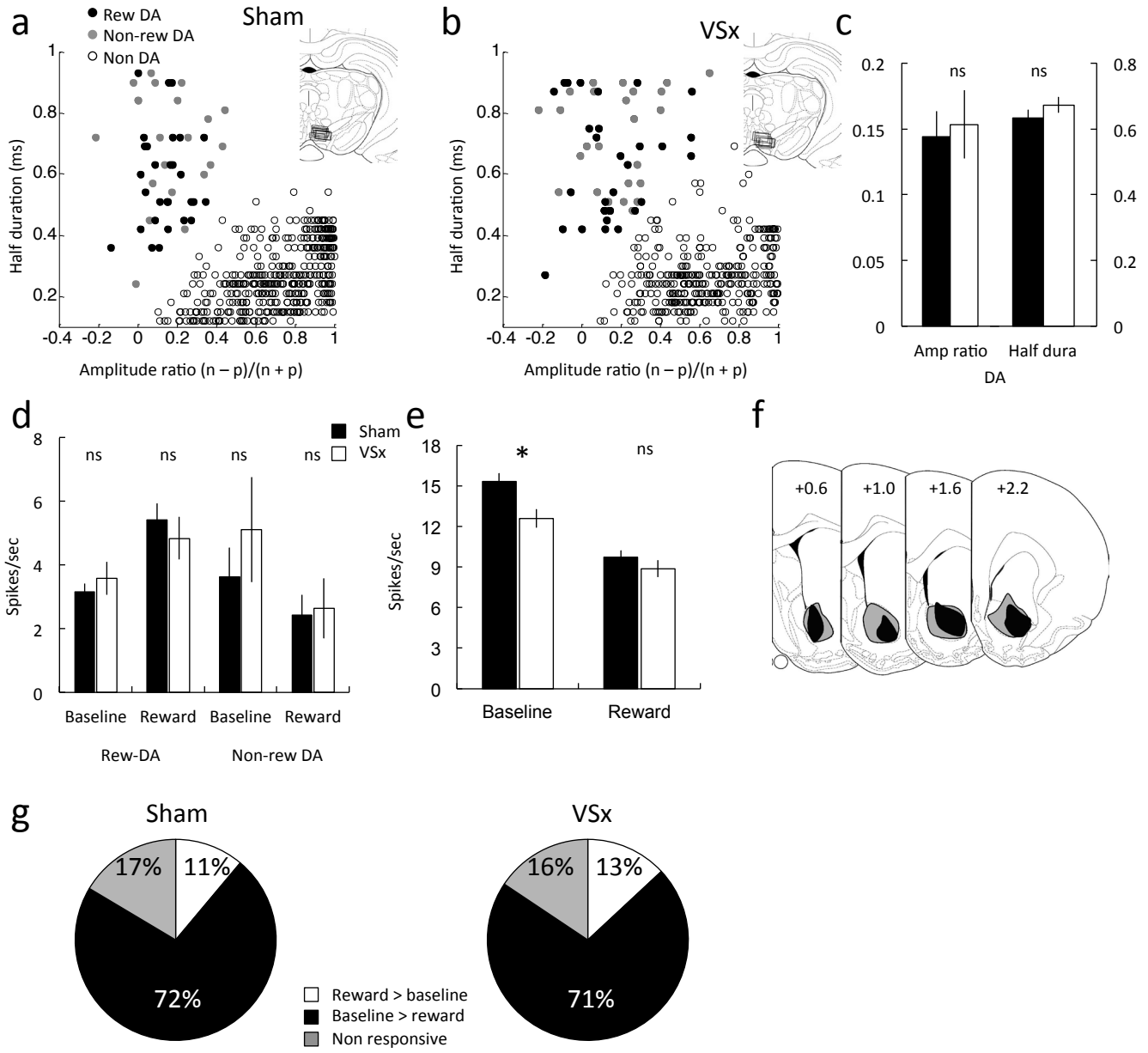

Figure S1, related to Figure 1: Identification, waveform features and firing rates of putative dopamine and non-dopamine neurons. (a, b) Results of cluster analysis based on the half time of the spike duration and the ratio comparing the amplitude of the first positive and negative waveform segments ((n – p)/(n + p)). The center and variance of each cluster was computed without data from the neuron of interest, and then that neuron was assigned to a cluster if it was within 3 s.d. of the cluster's center. Neurons that met this criterion for more than one cluster were not classified. This process was repeated for each neuron. Reward-responsive dopamine neurons (rew DA, n = 30 in sham, n = 31 in VSx), black; reward-nonresponsive dopamine neurons (non-rew DA, n = 21 in sham, n = 24 in VSx), gray; neurons that classified with other clusters, no clusters or more than one cluster (non DA, n = 450 in sham, n = 352 in VSx), open circles. Insets in each panel indicate location of the electrode tracks in sham (n = 9) (a) and VS-lesioned rats (n = 7) (b). **(c)** Bar graphs indicating average amplitude ratio and half duration of putative dopamine neurons in sham (black) and VS-lesioned rats (white). **(d)** Average baseline firing (left) and average firing to reward of reward-responsive (rew DA) and nonreward-responsive (non-rew DA) dopamine neurons in sham (black) and VS-lesioned (white) rats. **(e)** Average baseline firing (left) and average firing to reward (right) of non-dopamine neurons in sham (black) and VS-lesioned rats (white). *p < 0.01 or better. NS, nonsignificant (see main text). Error bars, s.e.m. **(f)** Brain sections illustrate the extent of the maximum (gray) and minimum (black) lesion at each level in VS in the lesioned rats (n = 7). **(g)** Pie-charts indicate populations of reward-responsive and non-responsive non-dopamine neurons in sham (left) and VS-lesioned rats (right). Reward-responsive neurons were classified by comparing firing between baseline and reward epoch (t-test, p < 0.05). A neuron showing a significantly higher firing to reward was represented as Reward > baseline, whereas a neuron showing a significant lower firing to reward was represented as Baseline > reward. Non-responsive was a neuron not showing significant difference in firing.

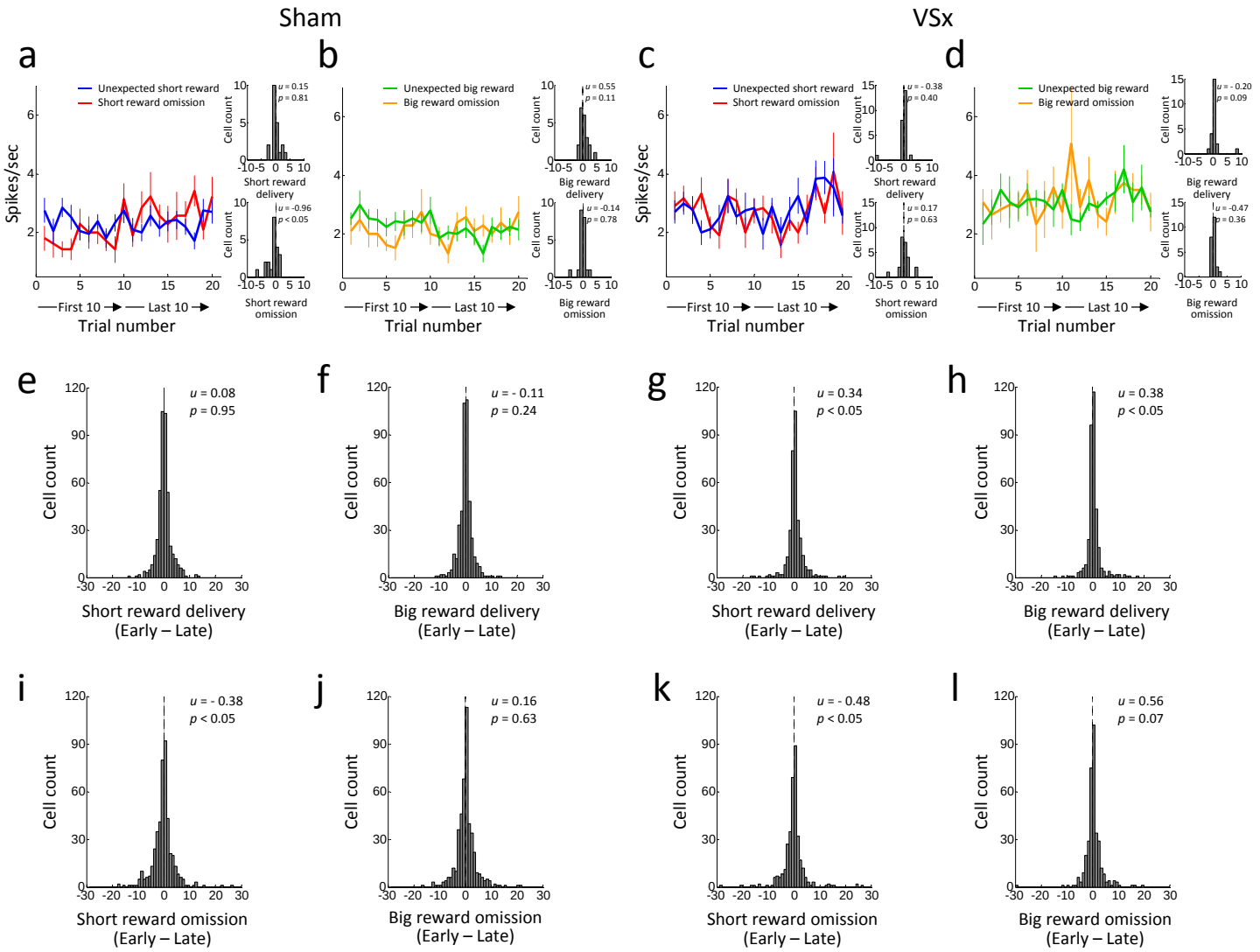

Figure S2, related to Figure 2: Changes in activity of reward non-responsive dopamine neurons and non-dopamine neurons to unexpected changes in timing and number of reward. (a – d) Average firing of reward non-responsive neurons (n = 21 in sham, n = 24 in VSx) during 500 ms after delivery of short reward (blue) and big reward (green), or omission of short reward (red) and big reward (orange) in sham (a and b) and VS-lesioned rats (c and d). Error bars, s.e.m. Small insets in each panel represent distribution of difference scores comparing firing to unexpected reward (top) and reward omission (bottom) early versus late in relevant trial blocks. **(e – h)** Distribution off difference scores in non-dopamine neurons (n = 450 in sham, n = 352 in VSx) comparing firing to unexpected short reward (e, g) and unexpected big reward (f, h) in sham (e, f) and VS-lesioned rats (g, h). **(i – l)** Distribution off difference scores comparing firing to omission of short reward (i, k) and omission of big reward (j, l) in sham (i, j) and VS-lesioned rats (k, l). Difference scores were computed from the average firing of each neuron in the first 5 and last 10 trials in relevant trial blocks. The numbers in upper right of each panel indicate results of Wilcoxon signed-rank test (p) and the average difference score (u).

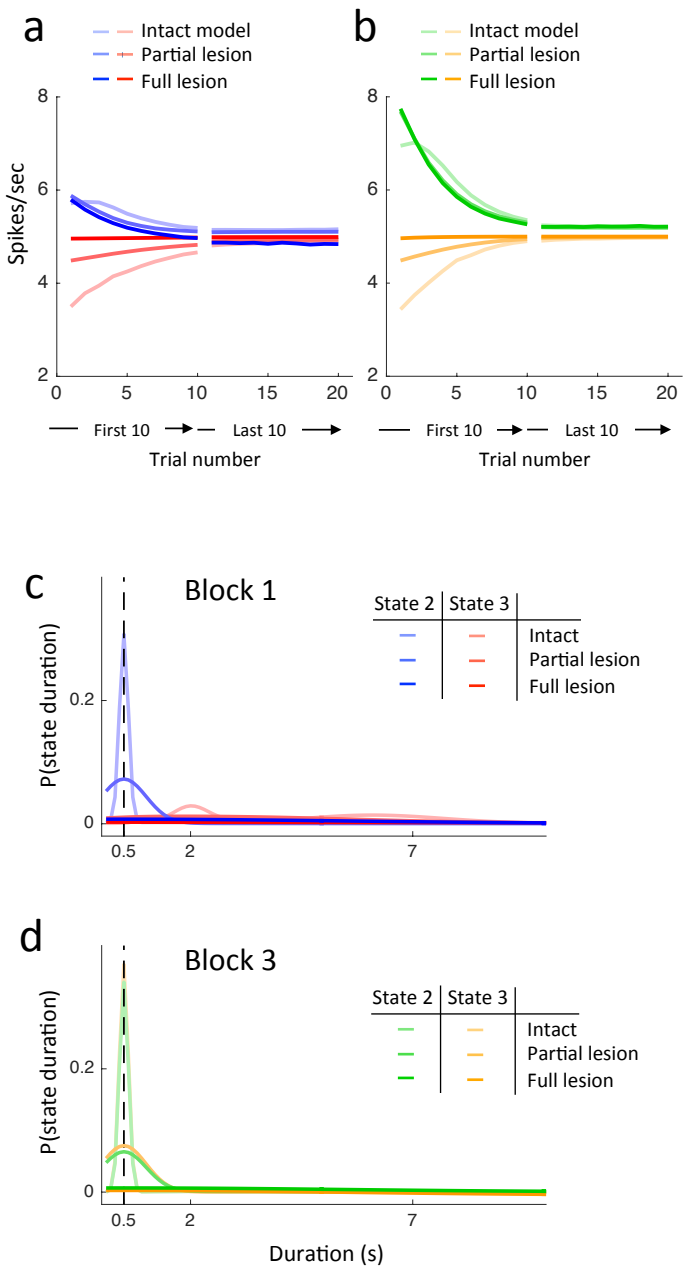

Figure S3, related to Figure 3: Effects of an alternative simulation of the lesion of temporal expectations in the semi-Markov TDRL model. Loss of temporal expectations was simulated by increasing the standard deviation of the kernel update for the learning of state dwell-time distributions, by a factor of 5 for a partial lesion and 50 for a full lesion. (a-b) Simulated average prediction errors during 500 ms after delivery of short reward (blue) and big reward (green), or omission of short reward (red) and big reward (orange) in the intact, partial and full lesion models. Note that the expectation of current state duration is not reduced in this lesion, and thus residual positive prediction errors to a reward delivered earlier than expected remain. (c) Learned dwell-time distributions at the end of block 1 (delay block) for state 2 in the short delay condition (blue) and state 3 in the long delay condition (red) for the intact, partial lesion and full lesion models in an example session. (d) Similar to (c) but at the end of block 3 (size block) for state 2 in the big reward condition (green) and state 3 in the small reward condition (orange). Any probability mass that falls below zero duration is reassigned to lie at infinite duration.

Category # of cells	reward-responsive DA n = 30	reward non-responsive DA n = 21	reward-responsive non DA n = 55	reward non-responsive non DA n = 395
met Criteria 1)-4)	23	0	6	8
PPE-NPE correlation	$p < 0.01$ $r = -0.57$	NA	$p = 0.75$ $r = -0.17$	$p = 0.39$ $r = -0.36$

Table S1, related to Figure 1: Single units meeting criteria for reward prediction error signaling by population in sham rats. Populations were identified as dopaminergic/non-dopaminergic and reward/non-reward-responsive as described in Figure S1. Neurons were classified as meeting criteria for prediction error signaling at the time of reward if they exhibited 1) increased firing to reward versus the empty well, 2) or omission, 3) a PPE index >1, and 4) a NPE index <1. Twenty three of 30 reward-responsive putative dopamine neurons described also in the main text passed this screen as error signaling, and they exhibited a significant and strong correlation in their response to PPE and NPEs. None of the other populations contained more than a few neurons that passed the criteria and firing to PPE and NPE's in these neurons was not correlated.

EXPERIMENTAL PROCEDURES

Computational modeling: We simulated learning and prediction error signaling in the task using temporal difference reinforcement learning in a semi-Markov framework with partial observability (Daw et al., 2006). Briefly, in this approach, we assume the rats represent the behavioral task as a sequence of states that each have an associated value, V, and a distribution over dwell times in that state, D. Observations during the task, such as an odor cue or the delivery of a reward in a well, signal a transition between states, at which time a prediction error is signaled and used to update state values. Additionally, transitions can occur without an external observation due to the mere passage of time (e.g., at the time that a reward was expected but failed to arrive, see below). That is, knowledge of the likely dwell time in a state (represented by D) can be used to infer a silent transition, and gate the signaling of a prediction error and the update of state values. To simulate a VS lesion in the model, we prevented the model from learning accurate dwell time distributions for each state, thereby degrading the ability of the model to infer these silent transitions when a reward is omitted or delayed. We now describe the model in more detail.

In accordance with the true task structure, we modeled the task using seven hidden states (s = 1,…,7; Fig. 3e). Formally, the task representation comprises also of an observation matrix O that gives the conditional probability of each observation given the underlying state, and a transition matrix T that defines the transition probabilities between states. We assumed that these two matrices are known to the rats, through their extensive training on the task. Figure 3e shows the model task representation with the possible transitions between the seven states marked by arrows, and the observation matrix O and transition matrix T. Observations included trial start (signaled by light on in the task), odor cues signaling the left and right wells, the first and second rewards, trial end, and a null (i.e., empty) observation. Following the true task structure, we assumed that each state gives rise to a single characteristic (non-empty) observation (though more than one state may have the same characteristic observation) that indicates a transition into that state. For example, observing the odor cue that signals the left well while in state 1 indicates a transition to state 3. Subsequent observation of a reward in the left well indicates entry to state 5, the only possible successor state from state 3. Formally, this means that the observation matrix O has high probability (here, 0.7) for only one non-null observation for each hidden state, apart from the null observation, which is associated equally with all states (p=0.2, with the remaining possible observations each assigned a background probability of 0.02). In the transition matrix T, we included a some small (10^{-6}) background transition probability between any two states, and apportioned the remaining probability equally amongst the possible transitions from each state (as marked in Fig. 3c). For simplicity, we treated free-choice trials to a certain side similarly to forced trials to that same side. Fully modeling the free choice cue and the choice behavior did not change any of the reported results.

"Partial observability" of states implies that states are not directly observable and are only probabilistically related to observations (as per the *O* matrix above), and that transitions between the states can also occur silently (i.e. with an empty observation), in which case the state transition must be inferred. This inference entails computing at every timepoint *t*+1 the probability of transitioning from state *s* at the previous timestep *t*, given all observations *o* until now, which we denote $\beta_{s,t+1}$:

$$\beta_{s,t+1} = P(s_t = s, \phi_t = 1 | o_1, \ldots, o_{t+1}).$$

Here, ϕ_t is a flag that indicates whether or not a state transition occurred between time *t* and *t*+1. Using Bayes' theorem,

$$\beta_{s,t+1} = \frac{P(o_{t+1}|s_t = s, \phi_t = 1) \cdot P(s_t = s, \phi_t = 1 | o_1, \ldots, o_t)}{P(o_{t+1}|o_1, \ldots, o_t)}$$

The first term in the numerator is the probability of the current observation given that the animal just transitioned from state *s* on the previous time point, which can be computed directly using the observation matrix *O* and transition matrix *T* by integrating over the successor state at time *t*+1: $\sum_{s'} T_{s,s'} O_{s',t+1}$. To calculate the second term in the numerator (which we denote $\alpha_{s,t}$), it is necessary to integrate over the possible durations *d* of the stay in state *s* since the time of the last non-empty observation t_o:

$$\alpha_{s,t} = \sum_{d=1}^{t_o} \frac{O_{s,o_{t-d+1}} D_{s,d} P(s_{t-d+1} = s, \phi_{t-d} = 1 | o_1, \ldots, o_{t-d})}{P(o_{t-d+1}, \ldots, o_t | o_1, \ldots, o_{t-d})},$$

where $D_{s,d}$ is the probability of dwelling *d* time in state *s*, which we assume the animals learn through experience by updating the estimated dwell times trial-by-trial as the task changed from block to block (details below), and $O_{s,o_{t-d+1}}$ is the probability of the observation made on entry to state *s* at the start of the duration *d*. Critically, computing this term relies on D_s, the (learned) distribution of dwell times in state *s*, such that a strong temporal expectation of a transition from state *s* after duration *d* will increase $\alpha_{s,t}$ (and consequently increase $\beta_{s,t+1}$) even in the absence of a concrete observation indicating state transition. This term is thus necessary for tracking the evolving expectation of a state transition based only on the passage of time, which is essential for inferring a state transition in the case of an omitted (or late) reward. Finally, the denominator in both equations is a normalization term that can be computed recursively by integrating the terms in the numerator over all possible states at times *t* (for $\beta_{s,t+1}$) or *t-d* (for $\alpha_{s,t}$). A more detailed derivation of the components of $\beta_{s,t+1}$ can be found in Daw et al. (2006).

Prediction errors for each state were signaled on each time point *t*+1 according to

$$\delta_{s,t+1} = \beta_{s,t+1}\left(e^{-\tau E[d_{s,t}]}\, r_{t+1} + \; e^{-\tau E[d_{s,t}]} E\left[\hat{V}_{s_{t+1}}\right] - \; \hat{V}_{s_t}\right),$$

where the term in brackets is analogous to the TD prediction error for fully-observable semi-Markov decision problems (Bradtke and Duff, 1995), and the $\beta_{s,t+1}$ term, which tracks the probability of a state transition, gates the prediction error. This gating by $\beta_{s,t+1}$ means that prediction errors are maximally signaled *only* at the (inferred) time of state transition. The prediction error term itself (i.e. the expression within brackets) describes exponential discounting of both rewards and the value of future states based on an estimate of how long the animal has been in the current state *s*. Here, r_{t+1} is the reward at the current timepoint, $E\left[\hat{V}_{s_{t+1}}\right]$ is the expected value of the next state, τ is a discount factor (set to 0.002) and $E\left[d_{s,t}\right]$ is the expected current dwell time in state *s*, used to discount the reward and value of this state. Since state transitions are not fully observable, the dwell time in any state is not known with certainty, therefore we use $E\left[d_{s,t}\right]$, computed by weighting all possible dwell times by their probability. Note that the maximal dwell time in the current state is the time that has passed since the last non-empty observation, t_o, thus

$$E\left[d_{s,t}\right] = \sum_{d=1}^{t_o} d \cdot P(d_t = d | s_t = s, \phi_t = 1, o_1, \ldots, o_{t+1}).$$

In the state-specific prediction error above, future rewards and states are therefore exponentially discounted according to an *estimate* of how long the animal has been in the current state. The total prediction error at time *t*+1 is the sum of the prediction errors for all states at this timepoint,

$$\delta_{t+1} = \sum\nolimits_s \delta_{s,t+1}$$

Each state value is updated on each time point *t*+1 using the total prediction error according to

$$V_s \leftarrow V_s + \; \eta E_{s,t+1} \delta_{t+1}$$

where $0 < \eta < 1$ is the learning rate or step size parameter and $E_{s,t+1} = \max_{t'}\left[\beta_{s,t'}\right]_{\tilde{t}}^{t+1}$ is the eligibility trace for state *s*. The eligibility trace tracks the probability that a state has been transitioned through at some point since the start of the trial (i.e. since time $\tilde{t}$), thus allowing updating of all states preceding (and thus predictive of) the current state, similar to the TD (1) algorithm in the standard TDRL framework (Sutton and Barto, 1998).

To learn the dwell-time distribution in each state, D_s was updated at the time of each non-empty observation using an iterative Gaussian-kernel density-estimation procedure. The update rule is $D_s \leftarrow D_s + \eta_D(K_d - D_s)$, where K_d is a Gaussian kernel, $N(d, CV\text{x}d)$, centered on the estimate of the time since the last observation d, and η_D is the dwell time distribution learning rate (set to 0.1). To account for known properties of scalar timing (Gibbon, 1977), the kernel was assumed to have a standard deviation proportional to the estimated duration d, with a fixed coefficient of variation $CV = 0.2$ (in accordance with empirical measurements of scalar timing noise, e.g. Gallistel et al., 2004). To ensure non-vanishing probabilities for all reasonable dwell times, we fixed a baseline probability of 10^{-4} for each time point. Using this learning rule, the dwell time distribution for each state asymptotically approached a Gaussian distribution centered on the mean duration of that state (based on timing of non-empty observations during the task).

Given their extensive training on the task, we assumed the animals began each session with an average expectation for both value and expected dwell time in each state. Accordingly, mean initial value for states 1, 2 and 3 was 0.7 (the average value of these states over the first two blocks) and all other states were initialized to mean zero. On each simulated session, values for each state were randomly initialized from a normal distribution with these state-specific mean values and a standard deviation of 0.005. The learning rate was set to $\eta = 0.3$ to ensure asymptotic values for each state were reached by the end of each block. Dwell-time distributions for the odor-cued states were initialized before each run with a Gaussian centered at 0.75s (the mean of the reward delay in the two wells at the start of a session) and with a standard deviation of $0.75\text{s}\times CV$. Dwell-time distributions for all other states were initialized to exponential distributions with a timescale of 10s. Simulations then used the actual task event timing from each neural recording session. As we did not explicitly model free-choice behavior, we replaced the free-choice cue on choice trials with the equivalent forced-choice cue for the chosen well in the task event timing sequences. As mentioned above, modeling the free choice trials separately did not qualitatively change any of the results. All parameters and initializations were chosen manually in order to achieve a qualitative fit to the neural data, but model results were not sensitive to specific parameter choices, and parameters affected model behavior as would be expected in a simple RL model (for example, setting a lower learning rate affects the time to asymptotic value for each state). Therefore, we did not attempt to formally fit the parameters to neural or behavioral data. In particular, results were not sensitive to initialization of either value or dwell time distribution in each state, with equivalent results obtained by setting initial values for all states to zero and all initial dwell time distributions to a uniform distribution over the range 0 to 50s. The main parameters of interest for modeling timing-related firing rates were the probability of receiving a null observation in each state (first column in the observation matrix in Fig 3e) and the precision of the learned dwell-time distributions for each state (controlled by the size and shape of the kernel for the update of the state dwell-time distributions). Together, these two model components control the possible size of negative prediction error signals to an omitted reward, the first by allowing a state transition to be inferred without a concrete observation (i.e.

by ensuring partial observability) and the second by 'opening' the gate on prediction error signals according to strong temporal expectations about state duration alone.

To simulate a VS lesion in the model, we prevented the model from accurately learning the distribution of dwell times in each state during the task by reducing the amplitude of the Gaussian kernel K_d in the dwell-time update by multiplying it by a 'lesion' fraction between 0 and 1, and apportioning the remaining probability mass to infinite duration (Fig. 3f,g). A factor of 0 corresponds to a full lesion, in which the kernel update has zero probability mass at the current time, and thus the dwell-time distributions for each state remain at the uniform baseline uncertainty of 10^{-4} for all finite time points. For consistency, we also lesioned the ability of the model to track dwell time within a state by multiplying the estimation of the expected duration $E[d_{s,t}]$ by the same fraction (this latter part of the lesion is not critical to any of the results reported here; however, it makes predictions for correlates of discounting, e.g., reaction times for differently delayed rewards, which can be tested in future work). This 'lesion' therefore blocked the formation of precise temporal expectations regarding the duration of task states by acting on the two key terms in the model that involve integration over durations, $\alpha_{s,t}$ and $E[d_{s,t}]$. To simulate a partial lesion, we set the lesion fraction to 0.5. Importantly, all other parameters in the model were left unchanged for both the partial and full simulated lesions.

Finally, to transform the simulated total prediction error into equivalent firing rates, we averaged over δ_{t+1} in the same 500ms epochs as used for the neural data analysis and rescaled these signals to neural firing rate as follows:

$$neural\ firing\ = baseline + scale\ factor \times prediction\ error.$$

For all simulations, baseline firing was set to 5Hz and positive and negative scale factors were 15 and 30 respectively.

REFERENCES

Bradtke, S.J., and Duff, M.O. (1995). Reinforcement learning methods for continuous-time Markov decision problems. In Advances in Neural Information Processing Systems, G. Tesauro, D.S. Touretzky, and T.K. Leen, eds. (Cambridge, MA: MIT Press), pp. 393-400.

Daw, N., Courville, A.C., and Touretzky, D.S. (2006). Representation and timing in theories of the dopamine system. Neural Computation 18, 1637-1677.

Gallistel, C.R., King, A., and McDonald, R. (2004). Sources of variability and systematic error in mouse timing behavior. Journal of Experimental Psychology: Animal Behavior Processes 30, 3-16.

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. Psychological Review 84, 279-325.

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