nature neuroscience

Dopamine reward prediction errors reflect hidden-state inference across time

Clara Kwon Starkweather[1], Benedicte M Babayan[1,2], Naoshige Uchida[1] & Samuel J Gershman[2]

Midbrain dopamine neurons signal reward prediction error (RPE), or actual minus expected reward. The temporal difference (TD) learning model has been a cornerstone in understanding how dopamine RPEs could drive associative learning. Classically, TD learning imparts value to features that serially track elapsed time relative to observable stimuli. In the real world, however, sensory stimuli provide ambiguous information about the hidden state of the environment, leading to the proposal that TD learning might instead compute a value signal based on an inferred distribution of hidden states (a 'belief state'). Here we asked whether dopaminergic signaling supports a TD learning framework that operates over hidden states. We found that dopamine signaling showed a notable difference between two tasks that differed only with respect to whether reward was delivered in a deterministic manner. Our results favor an associative learning rule that combines cached values with hidden-state inference.

Midbrain dopamine neurons are thought to drive associative learning by signaling the difference between actual and expected reward, a process termed reward prediction error (RPE)[1–4]. In particular, dopaminergic responses bear a striking resemblance to the error signal in a simple machine-learning algorithm known as temporal difference (TD) learning[1,5]. Several observations support this hypothesis[1–3]. Unexpected reward delivery elicits a large phasic burst of spikes from dopamine neurons. After an animal learns that a sensory cue predicts reward, dopamine neurons burst following the reward-predictive cue, and their phasic response is reduced following reward delivery. If a predicted reward is omitted, then dopamine neurons pause at the time when the animal usually receives reward.

Some of the theoretical assumptions in the original TD model are not realistic. For one, the TD learning model assumes that the agent assigns values to 'states'—representations of environmental conditions at any given time, which are classically specified in terms of observable stimuli. However, in the real world, stimuli often provide ambiguous information about states; the true underlying states are 'hidden' and must, therefore, be inferred[6,7]. For example, a lion crouching in the savannah might be indistinguishable from the tall grass, but these two objects carry very different consequences for an antelope. A principled way to incorporate hidden states into the TD learning framework is to replace the traditional stimulus representation with a belief state, which tracks the probability of being in each state given the trial history. This revised TD framework generates a value prediction that is computed on an inferred belief state. Although this idea has been explored theoretically[8,9], the empirical evidence remains sparse.

Here we designed two tasks to test whether dopaminergic RPEs provide evidence for a value prediction computed on a belief state. In both tasks, the cue–reward interval (ISI) was varied across trials. In the first task, the reward was delivered deterministically (100% rewarded). Our first task resembled other studies that examined dopamine signaling in tasks with variable ISIs[10–12]. This previous work, particularly Fiorillo *et al.*[10], described a mathematical framework for how temporal expectation influences dopamine RPEs. These studies demonstrated that a hazard function, or a temporally blurred 'subjective' hazard function, describes temporal expectancy in the case of 100% reward delivery. Expanding on this previous work, we also included a second task in which reward was occasionally omitted (90% rewarded). In the second task, the design was such that the animal could not initially be sure of whether the absence of reward meant that it was delayed or omitted entirely. As time elapsed following cue onset, the animal's belief that the reward would arrive would gradually yield to the belief that an omission trial occurred. Our results showed notable differences in dopamine signaling in these two tasks, which can be accounted for by incorporating hidden-state inference into the value prediction generated by the TD model. These results provide new evidence that dopaminergic RPEs are shaped by state uncertainty.

RESULTS

Behavioral task and electrophysiology

We trained mice on either of two tasks (**Fig. 1a**,**b**) (separate sets of three and four mice in tasks 1 and 2, respectively). In task 1, reward-predicting odors A–C forecasted reward delivery in 100% of trials. In task 2, odors A–C forecasted reward delivery in 90% of trials. In both tasks, the ISI following odor A was drawn from a discretized Gaussian distribution (2.0 ± 0.5 s, mean ± s.d.) defined over nine time points ranging from 1.2–2.8 s (**Fig 1c**,**d**). Trials for odors B and C had constant ISIs of 1.2 s and 2.8 s, respectively. We included odors B and C to examine the effect of temporal delay on dopamine RPEs. In the trials for odor D, reward was never delivered. In a subset of the mice that were trained on task 2 (**Supplementary Fig. 1a**; task 2b), we included

[1]Center for Brain Science, Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, USA. [2]Center for Brain Science, Department of Psychology, Harvard University, Cambridge, Massachusetts, USA. Correspondence should be addressed to N.U. (uchida@mcb.harvard.edu) or S.J.G. (gershman@fas.harvard.edu).

Received 17 July 2016; accepted 25 January 2017; published online 6 March 2017; doi:10.1038/nn.4520

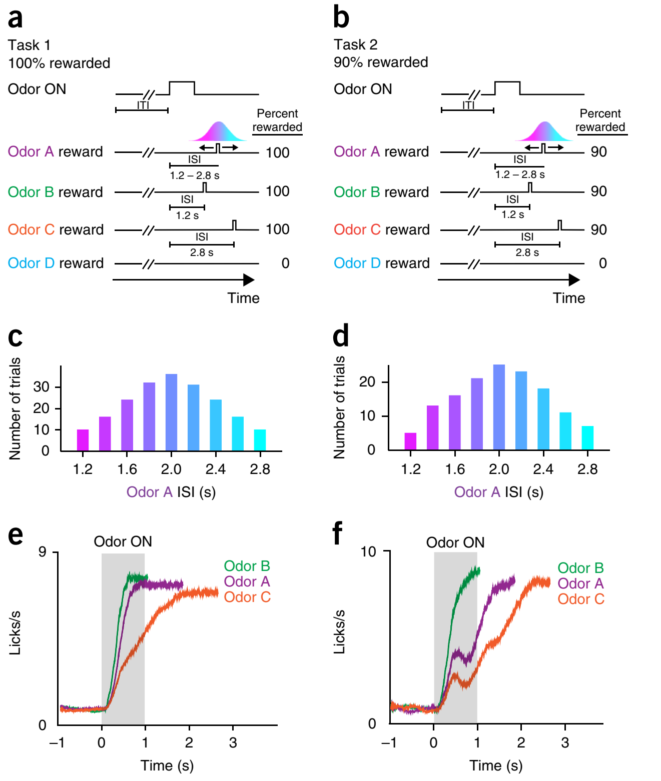

Figure 1 Task design. (**a**) In task 1, rewarded odors forecasted a 100% chance of reward delivery. Trials for odors B and C had constant ISIs, whereas trials for odor A had a variable ISI that was drawn from a discretized Gaussian distribution defined over nine time points ($n = 3$ mice). (**b**) In task 2, rewarded odors forecasted a 90% chance of reward delivery ($n = 4$ mice). ISIs for each odor were identical to those in task 1. (**c**) Histogram of ISIs for odor A trials during a sample task 1 recording session, showing nine possible reward delivery times. (**d**) Histogram of ISIs for odor A trials during a sample task 2 recording session. (**e**,**f**) Averaged non-normalized peristimulus time histogram (PSTH) for licking behavior across all task 1 ($n = 29$ sessions) (**e**) and task 2 ($n = 9$ sessions, not including task 2b) (**f**) recording sessions. Animals lick sooner for odor B (ISI = 1.2 s) than for odor C (ISI = 2.8 s). Licking patterns for odor A (variable ISI centered around 2.0 s) fall in between licking patterns for odor B and odor C.

an odor followed 2 s later by a reward to compare dopamine RPEs in omission trials for constant versus variable ISIs. Mice learned to lick in anticipation of a water reward following reward-predicting odors in tasks 1 and 2 (anticipatory licking in odor A–C ≠ baseline; $F_{1,50} > 150$, $P < 6 \times 10^{-17}$ for odors A–C in task 1, one-way analysis of variance (ANOVA); $F_{1,16} > 60$, $P < 1 \times 10^{-6}$ for odors A–C in task 2, one-way ANOVA; $F_{1,42} > 100$, $P < 5 \times 10^{-13}$ for odors A–C in task 2b, one-way ANOVA; **Supplementary Fig. 2a**–**c**). Mice ramped up their licking rates sooner and more steeply for odor B (ISI = 1.2 s) than for odor C (ISI = 2.8 s) (**Fig 1e**,**f**), and for 100% of rewarded odors than for 90% of rewarded odors (**Fig 1e**,**f** and **Supplementary Figs. 2d**,**e** and **3**). In both tasks 1 and 2, licking patterns in the trials for odor A (average ISI = 2.0 s) fell in between licking patterns observed in trials for odor B and odor C. Lick rates following odor D, which never predicted reward, did not change significantly from baseline rates ($F < 2.5$, $P > 0.10$ for both tasks, one-way ANOVA) (**Supplementary Fig. 2a**–**c**), demonstrating that mice learned the odor–outcome association.

We recorded the spiking activity of neurons in the ventral tegmental area (VTA) of the brain (387 neurons in seven mice; see **Supplementary Fig. 4** for recording sites) while the mice performed task 1 or task 2. To unambiguously identify dopamine neurons, we expressed the light-gated cation channel channelrhodopsin-2 (ChR2) in dopamine neurons. We delivered pulses of blue light through an optical fiber positioned near our electrodes and classified units as dopaminergic when they responded to light reliably and with a short latency period (**Supplementary Fig. 5** and Online Methods)[3,4,13].

Dopamine RPEs show opposing patterns of modulation across time in tasks 1 and 2

We recorded optogenetically identified dopamine neurons in tasks 1 and 2 (**Fig. 2**). In task 1 (100% reward probability), reward delivery elicited a phasic burst in dopamine firing ('post-reward firing') that was significantly modulated by ISI length (**Fig. 2a**). Post-reward firing was greatest for the shortest ISI and least for the longest ISIs (**Fig. 2a**). We quantified post-reward firing as the firing rate between 50 and 300 ms following water-valve opening, minus the firing rate 1,000 to 0 ms prior to odor onset. We quantified post-reward firing beginning at 50 ms after water-valve opening to distinguish between temporal modulation of pre-reward firing and post-reward firing, because the dopaminergic phasic response began 50 ms after valve opening (**Supplementary Fig. 6**). Furthermore, a recent study showed that intra-trial changes in dopamine firing may signal information distinct from pure RPEs, which prompted us to choose a single pre-cue baseline rather than an intra-trial baseline[14]. Our quantification of post-reward firing revealed that, on average, post-reward firing was modulated negatively by time (**Fig. 2g**). In addition to post-reward firing, we also found that the firing rate just prior to reward delivery ('pre-reward firing') was modulated over time (**Fig. 2g**). We computed pre-reward firing as the firing rate 400 to 0 ms prior to reward onset minus the firing rate 1,000 to 0 ms prior to odor onset. We found that pre-reward firing mirrored the post-reward pattern of negative modulation by time (**Fig. 2a**,**g**). Therefore, in the case of 100% reward probability, both pre-reward and post-reward dopamine firing decreased as a function of time. This result was consistent with other studies that have examined the effect of variable ISI length on dopaminergic RPEs in the case of 100% reward delivery (or reward-predicting event occurrence)[10–12].

We next explored how manipulating the certainty of reward delivery would alter the pattern of RPE modulation across time. In task 2, odor A's ISI was drawn out of the same Gaussian distribution as before, but reward was given in only 90% of trials. Pre-reward and post-reward firing in task 2 was calculated as that described for task 1. We found that post-reward firing was significantly modulated by ISI length (**Fig. 2b**). Notably, we observed the opposite trend of modulation over time, as compared to that in task 1. On average, reward delivery elicited a phasic response that was the least for the shortest ISI and greatest for the longest ISI (**Fig. 2b**,**h**). Pre-reward firing in task 2 was also significantly modulated by ISI length and tended to decrease throughout the variable ISI interval (**Fig 2b**,**h**). In summary, in the case of 90% reward probability, pre-reward firing decreased as a function of time, and post-reward firing increased as a function of time.

We asked whether these trends of temporal modulation could be seen at the level of individual neurons. In each task, and for each neuron, we plotted the post-reward firing rate versus the ISI on every trial and drew a best-fit line through the data (**Fig. 3a**–**f**). On the basis of the slopes of these best-fit lines, we found that 23/30 neurons tended toward negative modulation as a function of time in task 1 (95% confidence interval (CI) < 0 for 11/30 neurons; **Fig. 3g**). In task 2, we found that post-reward RPEs for 33/43 neurons tended to be positively modulated by time (95% CI > 0 for 14/43 neurons; **Fig. 3h**).

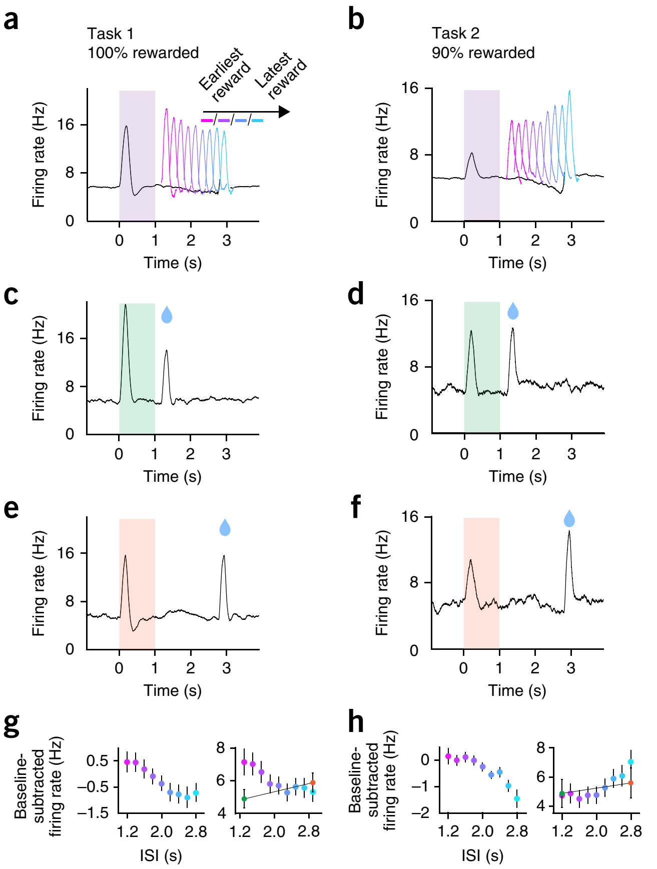

Figure 2 Averaged dopamine activity in tasks 1 and 2 shows different patterns of modulation over a range of ISIs. (**a**) Average non-normalized PSTH for all 30 dopamine neurons recorded during the trials for odor A in task 1. Shaded rectangle indicates 'odor ON'. Average pre-reward and post-reward dopamine RPEs were negatively modulated by time (post-reward firing: $F_{8,232} = 5.56$, $P = 1.9 \times 10^{-6}$, two-way ANOVA; factors: ISI, neuron; pre-reward firing: $F_{8,232}$=4.76, $P = 2.0 \times 10^{-5}$, two-way ANOVA; factors: ISI, neuron). (**b**) Average PSTH for all 43 dopamine neurons recorded during trials for odor A in task 2 (includes neurons from task 2b). Pre-reward dopamine RPEs (0–400 ms prior to reward onset) tended to be negatively modulated by time, whereas post-reward RPEs (50–300 ms following reward onset) tended to be positively modulated by time (post-reward firing: $F_{8,336} = 8.23$, $P = 3.48 \times 10^{-10}$, two-way ANOVA; factors: ISI, neuron; pre-reward firing: $F_{8,336} = 7.86$, $P = 1.0 \times 10^{-9}$, two-way ANOVA; factors: ISI, neuron). (**c**–**f**) Average PSTHs in trials for odor B (**c**,**d**) and odor C (**e**,**f**) in tasks 1 ($n = 30$ neurons) (**c**,**e**) and 2 ($n = 14$ neurons) (**d**,**f**). The blue droplet indicates reward delivery. (**g**,**h**) Summary plots for average pre-reward (left) and post-reward (right) firing in tasks 1 (**g**) and 2 (**h**). Green and orange circles represent averaged post-reward firing rates for odors B and C, respectively ($n = 30$ neurons in task 1; $n = 14$ neurons in task 2). Other colored circles represent averaged pre-reward and post-reward firing rates for odor A ($n = 30$ neurons in task 1; $n = 43$ neurons in task 2). In task 2, post-reward firing for odor B trials was not significantly different from post-reward firing for odor C trials (compare green and orange circle; $F_{1,13} = 0.85$, $P = 0.37$, two-way ANOVA; factors: odor, neuron). In task 2, post-reward firing for the latest possible ISI in the odor A trials was significantly different from the post-reward firing observed for odor C trials (compare orange and blue circles; $F_{1,13} = 7.15$, $P = 2 \times 10^{-2}$, two-way ANOVA; factors: odor, neuron) (data are mean ± s.e.m.).

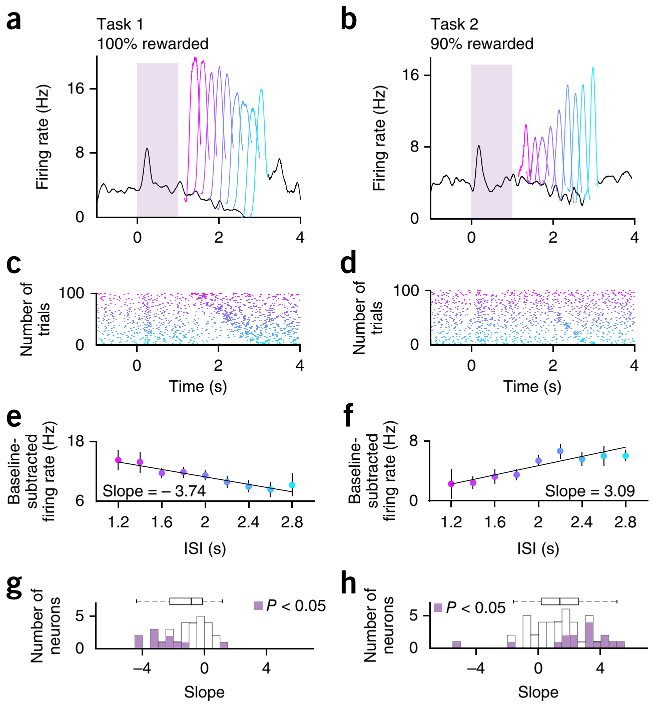

Figure 3 Individual dopamine neurons show opposing patterns of post-reward firing in tasks 1 and 2. (**a**,**b**) PSTH for two representative dopamine neurons during trials for odor A for a single recording session in task 1 (**a**) or task 2 (**b**), respectively. (**c**,**d**) Raster plots for the first 100 odor A trials of a single recording session in task 1 (**c**) or task 2 (**d**). (**e**,**f**) Examples of single-unit analysis. A best-fit line was drawn through a plot relating the ISI to the post-reward firing rate (50–300 ms following reward onset) for each odor A trial in task 1 ($n = 142$ odor A trials) (**e**) or task 2 ($n = 156$ odor A trials (data are mean ± s.e.m.) (**f**). (**g**,**h**) Slopes of best-fit lines in task 1 (**g**) or task 2 (**h**), as shown in **e**,**f**, for all dopamine neurons recorded ($n = 30$ neurons in task 1; $n = 43$ neurons in task 2). Bars to the right of "0" indicate neurons that tend to show positive temporal modulation of post-reward firing, whereas bars to the left of "0" indicate neurons that tend to show negative temporal modulation of post-reward firing. Purple shading indicates $P < 0.05$, or a 95% CI for the slope coefficient that does not overlap with 0. Box shows interquartile range (Q1–Q3, with median denoted in between); maximum whisker length is 1.5 × interquartile range.

We repeated the same analysis for pre-reward firing in both tasks. In task 1, pre-reward firing for 19/30 individual neurons tended toward negative modulation as a function of time (95% CI < 0 for 14/30 neurons). In task 2, pre-reward firing for 32/43 neurons tended toward negative modulation as a function of time (95% CI < 0 for 9/43 neurons). Therefore, individual neurons recorded in each task tended to reflect the trends of the pre-reward and post-reward temporal modulation described above.

To summarize, in tasks 1 and 2 we found that dopaminergic RPEs were modulated over time for various lengths of ISI. Pre-reward firing in both tasks tended to decline as a function of time. However, post-reward firing showed opposite trends of temporal modulation in these two tasks. In task 1, post-reward firing showed negative temporal modulation, whereas in task 2, post-reward firing showed positive temporal modulation.

Dopaminergic RPEs in task 2 cannot be explained by ISI length

Previous studies have demonstrated that phasic dopamine RPEs are sensitive to ISI length[10,15,16]. Specifically, post-reward firing is greater for longer ISIs, suggesting that growing temporal uncertainty increases the dopamine reward response. We asked whether the positive temporal modulation of post-reward firing in task 2 could be attributed to ISI length alone. If this were true, then we would expect the difference between post-reward firing in trials for odor B (ISI = 1.2 s) and those

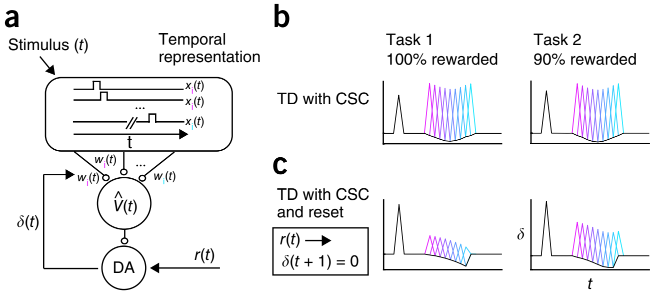

Figure 4 TD with CSC model, with or without reset, is inconsistent with our data. (**a**) Schematic of TD with CSC model adapted from ref. 1. The CSC temporal representation comprises features $x(t) = \{x_1(t), x_2(t), \ldots\}$ that are weighted to produce an estimated value signal $\hat{V}(t)$. $\delta(t)$ reports a mismatch between value predictions, and is used to update the weights of corresponding features. DA, dopamine; $r(t)$, reward at time t; $w_i(t)$, weight of feature i at time t. (**b**) TD with CSC produces a pattern of RPEs that resembles a flipped probability distribution, for both tasks 1 and 2. (**c**) TD with CSC and reset produces a pattern of RPEs that decreases over time, for both tasks 1 and 2. In **b**,**c**, plots are averaged from ten simulations of 5,000 trials each.

in trials for odor C (ISI = 2.8 s) to account for the difference in post-reward firing for the earliest and latest rewards in the trials for odor A (ISI = 1.2 s and 2.8 s, respectively; **Fig. 2h**). In task 2, we found that the average post-reward firing rate for odor C was about 1 Hz higher than that for odor B (**Fig. 2d**,**f**). This modest difference was not significant (**Fig. 2h**). Moreover, the latest possible reward delivery following odor A administration (ISI = 2.8 s) elicited post-reward firing that was significantly higher than that for odor C (**Fig. 2h**). These results indicated that the positive temporal modulation of post-reward firing observed in task 2 could not be attributed to ISI length alone.

TD learning with a complete serial compound representation cannot explain dopamine RPEs in tasks 1 and 2

Dopaminergic RPEs are believed to signal the error term in TD learning models[1]. We therefore examined whether previously proposed TD learning models could account for the dopamine signals observed in tasks 1 and 2.

In reinforcement-learning models, including TD learning models, value is typically defined as the expected discounted cumulative future reward[17]:

$$V(t)=E\left[\sum_{\tau=t}^{\infty}\gamma^{\tau-t}\,r(\tau)\right] \tag{1}$$

where $E[\cdot]$ denotes an average over randomness in reward delivery, and γ is a discount factor that down-weights future rewards. The goal of reinforcement-learning models is to learn correct-value estimates to maximize future rewards.

The original application of TD learning to the dopamine system[1] assumed a 'complete serial compound' (CSC) representation $x(t) = \{x_1(t), x_2(t), \ldots\}$ as stimulus features for value computation (**Fig. 4a**). The onset of a reward-predictive stimulus initiates a ballistic sequence of sub-states that mark small post-stimulus time steps. At a given time after the stimulus, only one of the sub-states $x_i(t)$ becomes active. In other words, $x_i(t) = 1$ exactly i time steps following stimulus onset, and $x_i(t) = 0$ in other time steps. The value function estimate is modeled as a linear combination of stimulus features:

$$\hat{V}(t)=\sum_i w_i x_i(t) \tag{2}$$

where w_i is a predictive weight associated with feature i. The weights are updated according to the following learning rule:

$$\Delta w_i=\alpha x_i(t)\,\delta(t) \tag{3}$$

where α is a learning rate and $\delta(t)$ is the RPE, computed according to:

$$\delta(t)=r(t)+\gamma\hat{V}(t+1)-\hat{V}(t) \tag{4}$$

We first tested whether TD learning with the CSC could explain our experimental results.

For both tasks 1 and 2, we found that TD learning with the CSC produced RPEs that were most suppressed for rewards delivered at the center of the Gaussian ISI distribution and least suppressed for rewards delivered at the tails of the distribution (**Fig. 4b**). The pattern of RPEs across different ISIs resembled a flipped distribution of experienced ISIs. Moreover, the modulation of RPEs across ISIs was identical between tasks 1 and 2, indicating that this model could not explain our data.

We next asked whether a simple modification of the original model, a 'reset' feature that sets the RPE to 0 after reward arrives[18,19], could better account for our results. This model rectifies one key inconsistency between the data and a simple CSC TD model: when a reward is delivered unexpectedly early, the 'pause' predicted by the CSC TD model at the usual time of reward does not occur[20]. When we trained a TD model with the CSC and reset on our tasks, the model produced a pattern of RPEs suggestive of a hazard function, i.e, reward expectation that grows over time, increasingly suppressing excitation toward the end of the variable ISI range (**Fig. 4c**). A pattern of decreasing RPEs over time matched our task 1 data. However, in task 2, this model also produced RPEs that generally decreased throughout the variable ISI range, deviating from the trend of our data. Therefore, this proposed modification to the original model could not explain our results. Our data did not completely rule out other reset devices, such as resetting the stimulus trace following reward[21]. However, as pointed out by Daw and colleagues[9], such a reset device assumes an inferred state change and may not generalize well to more complex scenarios with multiple rewards.

TD learning with belief states explains dopaminergic RPEs in tasks 1 and 2

Another way to approach these data is to reconsider the computational problem being solved by the animal. One potentially important problem for the mice in the above tasks is knowing whether they are in one of the two states—the ISI state, during which the animal expects a reward, or the ITI state, during which no reward is expected. In task 1, these two states were fully observable, as cue onset unambiguously signaled a transition to the ISI state and reward onset unambiguously signaled a transition to the ITI state; no transitions occurred without one of these events (**Supplementary Fig. 7a**). Thus, the states were fully observable, and the only computational problem was predicting reward. In task 2, omission trials caused the ITI state to self-transition (while still emitting a cue). This meant that both ITI-to-ISI and ITI-to-ITI transitions generated the same observation, rendering the states partially observable or 'hidden' (**Supplementary Fig. 7b**). Thus, task 2 introduces an additional computational problem, hidden-state inference.

In this framework, a critical computation is to assign a probability of being in the ITI or ISI state at a given moment. To incorporate this process in our model, here we assumed that the ISI and ITI comprise temporal 'sub-states', with the analogy to the CSC model (**Fig. 5a**,**b**). The normative solution to the hidden-state inference problem is given

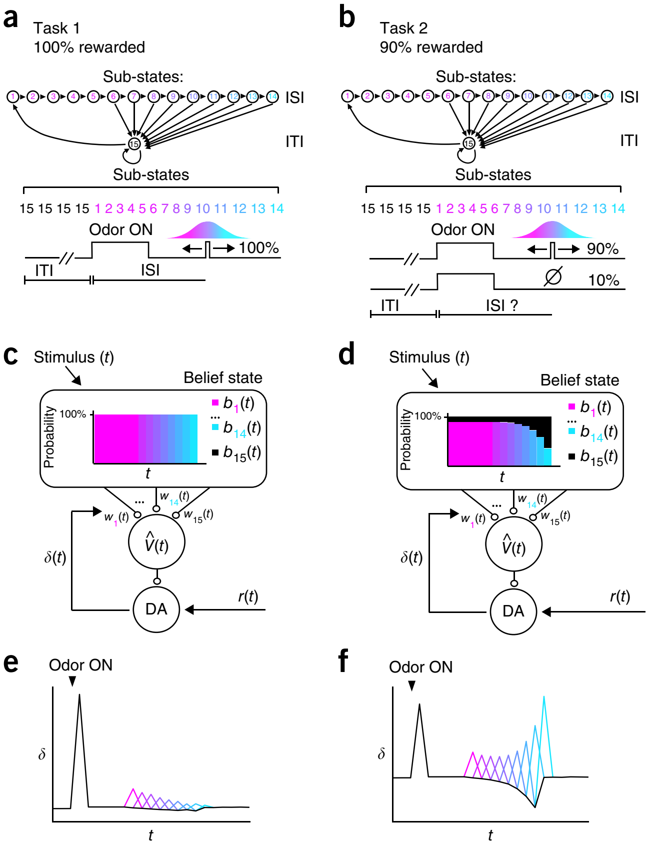

Figure 5 Belief-state model is consistent with our data. (**a**,**b**) Schematic for our model, in which the ISI and ITI states comprise sub-states 1–15 in task 1 (**a**) and task 2 (**b**). (**c**,**d**) The CSC temporal representation is swapped for a belief state. Expected value is the linear sum of both weight and belief state

$$\hat{V}(t)=\sum_i w_i b_i(t)\,.$$

In task 1 (**c**), the belief state sequentially assigns 100% probability to each ISI sub-state as time elapses after odor onset. In task 2 (**d**), the belief state gradually shifts in favor of the ITI as time elapses and reward fails to arrive. (**e**,**f**) Plots averaged from one representative simulation for task 1 (n = 3,000 simulated reward trials) (**e**) and task 2 (n = 2,700 simulated reward trials) (**f**) using the belief-state model (quantification in **Supplementary Fig. 8**). In addition to capturing the opposing post-reward firing patterns between task 1 and task 2, this model also captures negative temporal modulation of pre-reward firing in both tasks.

by Bayes' rule, which stipulates how an animal's probabilistic beliefs about states should be updated over time:

$$b_i(t+1)\propto p\big(o(t)\big|i\big)\sum_j p\big(i\big|j\big)b_j(t-1) \qquad (5)$$

where $b_i(t)$ is the posterior probability that the animal is in sub-state i at time t, $p(o(t)\,|i)$ is the likelihood of the observation $o(t)\in\{cue,reward,null\}$ under the hypothetical sub-state i, and $p(i|j)$ is the probability of transitioning from sub-state j to sub-state i (**Supplementary Fig. 7c–f**).

The vector $b(t)$ functions as a belief state that can substitute for the CSC in learning equations 2 and 3 shown above (**Fig. 5c**,**d**). In task 1, in which the observations were unambiguous, the belief state was identical to the CSC. In task 2, the belief state departed from the CSC by representing subjective uncertainty about the current sub-state (the posterior probability of being in the ISI or ITI state can be computed from this representation by summing the belief-state vector over all sub-states within a particular state).

Although we have formulated the model in terms of probabilities over sub-states, the model could have alternatively been formulated in continuous time using semi-Markov dynamics[9], in which sub-states are replaced by dwell-time distributions in each state. These models are mathematically equivalent; we chose the sub-state formulation to draw clearer connections to other models (a point to which we return in the Discussion).

A belief-state TD model produced error signals that resembled dopamine RPEs in our tasks (**Fig. 5e**,**f**). In task 1, the states were fully observable, and thus the belief state was uniform throughout the variable ISI range (**Figs. 5c** and **6a**,**b**)—as soon as the cue came on, the belief state encoded a 100% probability of being in one of the ISI sub-states and a 0% probability of being in the ITI sub-state. Because the momentary probability of receiving a reward was greater at later ISIs than at earlier ISIs, the sub-states at later ISIs accrued higher weights than the sub-states at earlier ISIs, producing a ramping value signal (**Fig. 6a**,**b**). This ramping value signal resulted in RPEs that were increasingly suppressed toward the end of the variable ISI range (**Figs. 5e** and **6a**,**b**; quantification in **Supplementary Fig. 8a**,**c**), producing a pattern of negative modulation by time similar to that observed in our task 1 data.

In task 2, the belief state takes into account the possibility of an unobservable state transition. Therefore, unlike in task 1, the belief state was not uniform throughout the variable ISI range. As time elapsed and the reward failed to arrive, the belief state progressively shifted in favor of the ITI state over the ISI state (**Figs. 5d** and **6d**). Rewards sometimes arrived at the latest ISIs, increasing the weights for the corresponding sub-states. However, the belief state for these late time points was so skewed toward the ITI state that the value signal actually decreased for the longer ISIs (**Fig. 6c**,**d**). This decreasing value signal resulted in pre-reward RPEs that were the most suppressed, and post-reward RPEs that were the least suppressed, at the end of the variable ISI range (**Figs. 5f** and **6c**,**d**; quantification in **Supplementary Fig. 8b**,**d**). Post-reward RPEs toward the end of the interval were nearly as large as the unpredicted rewards, both in the results from our model (**Supplementary Fig. 8d**) and in our data (**Supplementary Fig. 9**). Therefore, our model captures the pattern of pre-reward and post-reward RPEs in our task 2 data. Of note, the belief-state model captures the opposing trends of temporal modulation for post-reward dopamine RPEs in tasks 1 and 2.

One additional empirical result that we compared with our belief-state TD model was task 2b reward omission responses. For omission trials with odor A (2-s variable ISI), we found that the trough of the dip in dopamine firing occurred slightly later than the trough observed in omission trials with odor B (2-s constant ISI). This shift in the trough of the omission response was also reproduced by our belief-state TD model (**Supplementary Fig. 1b**,**c**).

One assumption of our model was that animals had perfectly learned the Gaussian distribution of ISIs. We lacked any behavioral indication that the animals had truly learned the probability distribution, so we tried relaxing this assumption in our model by instead training it on a uniform distribution of ISIs. We found that our model produced the same 'flip' in the temporal modulation of post-reward RPEs between tasks 1 and 2 when trained on a uniform distribution (**Supplementary Fig. 10**). Therefore, our modeling result is relatively agnostic to the precise shape of the learned ISI distribution.

Finally, although our model captured the trends of pre-reward and post-reward temporal modulation in both tasks 1 and 2, the overall

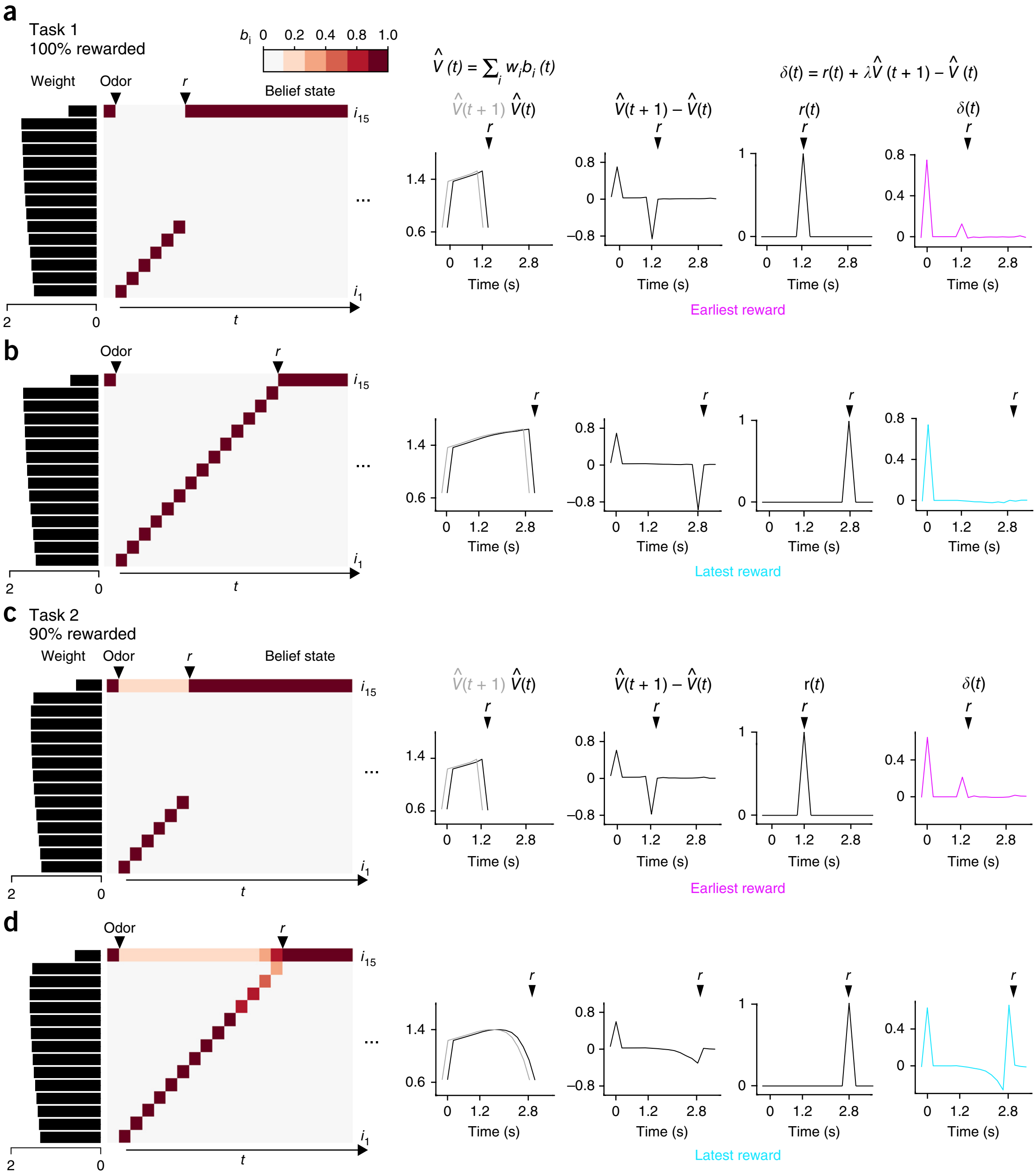

Figure 6 Value signals in the belief-state model of tasks 1 and 2. (**a**,**b**) Plots averaged from one representative simulation for task 1 ($n = 3{,}000$ simulated rewarded trials). Plots (left) are belief state and weights for earliest (**a**) and latest (**b**) rewards in task 1. Graphs (right) are for the corresponding value function, reward and RPE. As time elapses following odor onset in task 1, the belief state proceeds through ISI sub-states i_1–i_{14} by sequentially assigning a probability of 100% to each sub-state. Later ISI sub-states accrue greater weights. Estimated value is approximated as the dot product of belief state and weight, producing a ramping value signal that increasingly suppresses $\delta(t)$ for longer ISIs. (**c**,**d**) Plots averaged from one representative simulation for task 2 ($n = 2{,}700$ simulated rewarded trials). As time elapses following odor onset in task 2, the belief state comprises a probability distribution that gradually decreases for ISI sub-states i_1–i_{14} and gradually increases for the ITI sub-state i_{15}. This produces a value signal that declines for longer ISIs, resulting in the least suppression of $\delta(t)$ for the latest ISI.

dopamine firing in task 1 was much larger than that predicted by our model. We asked what the cause of the discrepancy in post-reward RPE magnitude between our task 1 data and model could be. Because our mice were trained on an odor–outcome association, the exact times when the animals sniffed and detected the odor ("odor ON") was jittered from trial to trial. This temporal jitter limits how precisely the animal can anticipate reward timing. Therefore, because our model did not incorporate this trial-by-trial jitter, it suppressed RPEs more effectively, particularly in task 1 conditions that allowed reward timing to be predicted perfectly by the end of the interval. Furthermore, our mice were trained for a relatively short length of time (~1–2 weeks) prior to recording, potentially limiting the extent to which RPEs could be suppressed. Indeed, training our model on fewer trials increased the magnitude of post-reward RPEs.

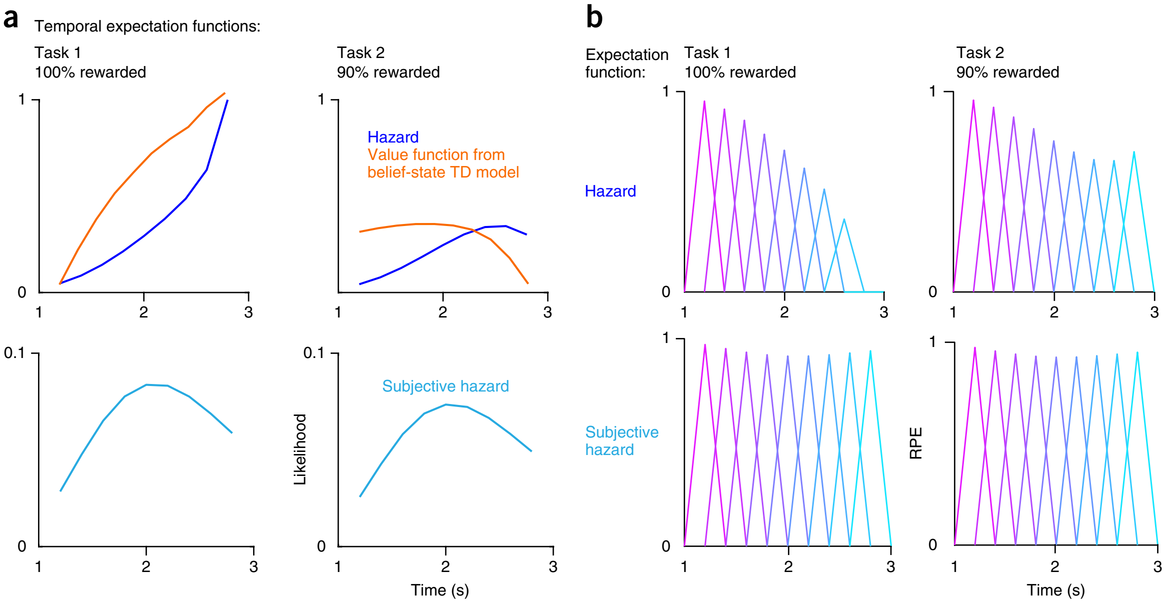

Figure 7 Hazard and subjective hazard functions for tasks 1 and 2. (**a**) Hazard (top) and subjective hazard (bottom) functions deviate substantially from the trend of value expectation over time in our belief-state TD model, particularly in task 2. Note the value functions are scaled versions of those shown in **Figure 6b**,**d** to aid visual comparison of trends over time. (**b**) Illustration of how RPEs would appear in our data if the reward expectation signal corresponded to hazard or subjective hazard functions.

Previous accounts of 'hazard-like' expectation signals cannot explain our data

Previous work has described hazard-like expectation signals that shape neural firing and animal behavior[22–25]. A hazard function is defined as the probability function divided by the survival function, or in other words, the likelihood that an event will occur given that it has not yet occurred. In studies that analyzed dopaminergic RPEs in tasks with variable ISIs[10–12], the variably timed event always occurred (100% event probability), and the ISI was drawn from a uniform distribution. With respect to both pre-reward and post-reward dopamine firing, all of these studies found a pattern of decreasing excitation over elapsed time, which was thought to correspond to a rising hazard function that increasingly suppressed RPEs at later ISIs. Furthermore, a functional magnetic resonance imaging study provided evidence that blood-oxygen-level-dependent (BOLD) signals in the VTA track hazard signals in humans[26]. However, one aspect of previous work could not be explained by using a hazard function—when animals were trained on an exponential distribution of ISIs, post-reward RPE's were still negatively modulated over time, despite the flat hazard function of the ISI distribution[10]. Notably, when we trained our belief-state TD model on an exponential distribution similar to that in this previous work, our model was able to reproduce the negative temporal modulation of post-reward RPEs (**Supplementary Fig. 11d**,**e**).

Our data in task 1, which used a Gaussian ISI distribution and 100% reward probability, also revealed a pattern of decreasing pre-reward and post-reward dopamine firing, which matched the proposal that a hazard function may describe the trend of temporal expectancy reflected by dopamine RPEs (**Fig. 7**). However, our data in task 2 could not be explained by a hazard function nor could they be explained by a temporally blurred subjective hazard function that was computed by blurring the probability distribution function with a Gaussian distribution whose s.d. scaled with elapsed time (refs. 23,24 and Online Methods) (**Fig. 7**). Plotting the hazard function and subjective hazard functions for task 2 revealed that both of these functions find a minimum for the earliest rewards (**Fig. 7a**). However, our data indicated that temporal expectation was at its maximum for the earliest rewards, because the earliest post-reward RPE's were most suppressed (**Fig. 2b**). We illustrated this contrast by plotting the value function from our belief-state TD model alongside the hazard function for task 2 (**Fig. 7a**). In summary, a hazard function may describe temporal expectancy for conditions in which rewards were given 100% of the time. However, temporal expectancy is dramatically altered in conditions involving uncertainty about whether the event will occur at all.

DISCUSSION

Here we examined how dopaminergic RPE signals change with respect to reward timing and probability. Our experimental results showed that, depending on whether or not a reward was delivered deterministically, dopaminergic RPEs exhibited opposite patterns of temporal modulation. Furthermore, our modeling result showed that these data are well explained by a TD model incorporating hidden-state inference[9]. Because dopaminergic RPEs are proposed to signal the error term in TD learning, these findings deepen our understanding of how TD learning may be implemented in the brain. TD learning uses RPEs to update the weights of task-related features, which are classically represented as a cascade of sub-states (CSC) that track elapsed time following stimulus onset[1,5]. Our findings support an alternative belief-state model that tracks a posterior distribution over sub-states.

A long-standing idea in modern neuroscience is that the brain computes inferences about the outside world rather than by passively observing its environment[27,28]. This is accomplished through the inversion of a generative model that maps hidden states to sensory observations. For example, the hidden state of a lion crouching in the grass could be mapped to sensory cues such as a faint rustling or a nearby paw print. By conditioning its belief state on observations of its environment, the antelope may predict the lion's presence. Following earlier theoretical work[8,9,29], we argue that this inferential process is at operational in the dopamine system. In particular, inferences about hidden states furnish the inputs into the reward-prediction machinery of the basal ganglia, with dopamine signaling errors in these reward predictions.

Our work follows two recent empirical studies that explored a state-based framework in the striatum and the VTA[30,31]. In the first of these studies, the authors found that individual striatal cholinergic interneurons preferentially fire for certain 'states', which mapped onto different blocks of a behavioral task[30]. In the second of these studies, a state-based model was used to capture the effect of a striatal lesion, which selectively affected the temporal specificity of dopaminergic prediction errors but spared value-related prediction errors[31]. These two studies support our claim that a belief-state representation may be operational in the basal ganglia reward-processing circuitry.

Previous studies have shown a pattern of decreasing RPEs over time during tasks in which ISIs are drawn from a uniform probability distribution[10–12]. We asked whether our model could account for the temporal modulation of dopamine RPEs in these previous studies. After training the belief-state TD model on a uniform distribution of reward timings, our model elicited negative temporal modulation of RPE signals (**Supplementary Fig. 11a,b**), indicating that our model is compatible with the data in these studies. However, we found that the belief-state TD model was not the only model that produced decreasing RPEs over time. TD learning using the CSC and reset also produced a pattern of decreasing excitation over time when trained on a flat probability distribution (**Supplementary Fig. 11c**). Because a 100%-rewarded condition fails to distinguish between these two models, it was critical that our experiments included both '100% rewarded' and '90% rewarded' tasks for comparison. Comparing RPEs for both of these task conditions allowed us to distinguish between the predictions of various associative learning models, thereby expanding on these previous studies.

The belief-state model provides a framework that is separate from, and entirely compatible with, previous work that examined the effect of temporal delay on dopamine RPEs[10,15,16]. These studies showed that dopamine RPEs are less suppressed for longer ISIs, likely due to scalar timing uncertainty. For simplicity, our belief-state model omitted the effect of temporal uncertainty to clearly demonstrate the effect of belief-state inference on the value function and dopamine RPEs. However, we can incorporate scalar temporal uncertainty into our model by blurring the belief-state distribution with a Gaussian kernel whose s.d. is proportional to elapsed time[31] (**Supplementary Fig. 12**). To create this 'blurred' belief-state model, we fit a scalar timing noise parameter to account for post-reward RPEs for 1.2-s and 2.8-s constant delays (odors B and C). This temporally blurred belief-state model still captured our data well in tasks 1 and 2.

Although we have focused on the belief-state TD model, another prominent model replaces the CSC with 'microstimulus' features—temporally diffuse versions of the discrete time markers in the CSC[32]. The microstimulus model incorporates neural timing noise that accrues for longer intervals by representing each sub-state's temporal receptive field as a Gaussian function whose s.d. increases, and amplitude decreases, with the post-stimulus interval. Although the microstimulus and belief-state models are typically thought of as alternatives[33], they can be conceived as realizations of the same idea at different levels of analysis.

By examining the belief state over time, we saw that the belief state over each sub-state peaks at a specific moment during the trial (**Fig. 6**). In task 2, the peaks became progressively lower as a function of time, due to the increased probability of a state transition. This decrease in amplitude mirrored the decrease in amplitude of microstimuli as a function of time. If we take into account noise and autocorrelation in neural signaling, then we expect these functions to become more temporally dispersed, further increasing the resemblance to microstimuli. This suggests that microstimuli might be viewed as a neural realization of the abstract-state representation implied by the belief-state model.

The key difference between microstimuli and belief states is that the shape of belief states is sensitive to task structure (e.g., the omission probability), whereas microstimuli have been traditionally viewed as fixed. However, if we view microstimuli as being derived from belief states, then we expect the microstimulus shape to change accordingly. Indeed, evidence suggests that microstimulus-like representations adapt to the distribution of ISIs, 'stretching' to accommodate distributions with a wider range of ISIs[34]. This is precisely what we would expect to see if the transition function in the belief-state model is adapted to the ISI distribution.

In summary, our data provide support for a TD learning model that operates over belief states, consistent with the general idea that the cortex computes probability distributions over hidden states that get fed into the dopamine system. Although belief states are cognitive abstractions, they could be realized in the brain by neurons with temporal-receptive fields resembling microstimuli.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS
We thank R. Born, J. Assad and C. Harvey for discussions, members of the Uchida lab for discussions and C. Dulac for sharing resources. This work was supported by National Science Foundation grant CRCNS 1207833 (S.J.G.), US National Institutes of Health grants R01MH095953 (N.U.), R01MH101207 (N.U.), T32 MH020017 (C.K.S.) and T32GM007753 (C.K.S.), a Harvard Brain Science Initiative Seed grant (N.U.), a Harvard Mind Brain and Behavior faculty grant (S.J.G. and N.U.) and Fondation pour la Recherche Medicale grant SPE20150331860 (B.M.B.).

AUTHOR CONTRIBUTIONS

C.K.S. and N.U. designed the recording experiments and behavioral task; C.K.S. collected and analyzed data; C.K.S. and S.J.G. constructed the computational models; and C.K.S., B.M.B., S.J.G. and N.U. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Schultz, W., Dayan, P. & Montague, P.R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
2. Bayer, H.M. & Glimcher, P.W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141 (2005).
3. Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B. & Uchida, N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88 (2012).
4. Eshel, N. *et al.* Arithmetic and local circuitry underlying dopamine prediction errors. *Nature* **525**, 243–246 (2015).
5. Sutton, R.S. & Barto, A.G. in *Learning and Computational Neuroscience: Foundations of Adaptive Networks* (eds. Gabriel, M. and Moore, J.) 497–537 (MIT Press, 1991).
6. Gershman, S.J., Blei, D.M. & Niv, Y. Context, learning and extinction. *Psychol. Rev.* **117**, 197–209 (2010).
7. Gershman, S.J., Norman, K.A. & Niv, Y. Discovering latent causes in reinforcement learning. *Curr. Opin. Behav. Sci* **5**, 43–50 (2015).
8. Rao, R.P.N. Decision making under uncertainty: a neural model based on partially observable Markov decision processes. *Front. Comput. Neurosci.* **4**, 146 (2010).
9. Daw, N.D., Courville, A.C. & Touretzky, D.S. Representation and timing in theories of the dopamine system. *Neural Comput.* **18**, 1637–1677 (2006).

10. Fiorillo, C.D., Newsome, W.T. & Schultz, W. The temporal precision of reward prediction in dopamine neurons. *Nat. Neurosci.* **11**, 966–973 (2008).
11. Pasquereau, B. & Turner, R.S. Dopamine neurons encode errors in predicting movement trigger occurrence. *J. Neurophysiol.* **113**, 1110–1123 (2015).
12. Nomoto, K., Schultz, W., Watanabe, T. & Sakagami, M. Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. *J. Neurosci.* **30**, 10692–10702 (2010).
13. Tian, J. & Uchida, N. Habenula lesions reveal that multiple mechanisms underlie dopamine prediction errors and prediction-error-based learning. *Neuron* **87**, 1304–1316 (2015).
14. Hamid, A.A. *et al.* Mesolimbic dopamine signals the value of work. *Nat. Neurosci.* **19**, 117–126 (2016).
15. Kobayashi, S. & Schultz, W. Influence of reward delays on responses of dopamine neurons. *J. Neurosci.* **28**, 7837–7846 (2008).
16. Jo, Y.S. & Mizumori, S.J.Y. Prefrontal regulation of neuronal activity in the ventral tegmental area. *Cereb. Cortex* **26**, 4057–4068 (2016).
17. Sutton, R.S. Learning to predict by the methods of temporal differences. *Mach. Learn.* **3**, 9–44 (1988).
18. Suri, R.E. & Schultz, W. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience* **91**, 871–890 (1999).
19. Suri, R.E. & Schultz, W. Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Exp. Brain Res.* **121**, 350–354 (1998).
20. Hollerman, J.R. & Schultz, W. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* **1**, 304–309 (1998).
21. Brown, J., Bullock, D. & Grossberg, S. How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *J. Neurosci.* **19**, 10502–10511 (1999).
22. Oswal, A., Ogden, M. & Carpenter, R.H.S. The time course of stimulus expectation in a saccadic decision task. *J. Neurophysiol.* **97**, 2722–2730 (2007).
23. Janssen, P. & Shadlen, M.N. A representation of the hazard rate of elapsed time in macaque area LIP. *Nat. Neurosci.* **8**, 234–241 (2005).
24. Tsunoda, Y. & Kakei, S. Reaction-time changes with the hazard rate for a behaviorally relevant event when monkeys perform a delayed wrist-movement task. *Neurosci. Lett.* **433**, 152–157 (2008).
25. Ghose, G.M. & Maunsell, J.H. Attentional modulation in visual cortex depends on task timing. *Nature* **419**, 616–620 (2002).
26. Klein-Flügge, M.C., Hunt, L.T., Bach, D.R., Dolan, R.J. & Behrens, T.E. Dissociable reward and timing signals in human midbrain and ventral striatum. *Neuron* **72**, 654–664 (2011).
27. Friston, K. A theory of cortical responses. *Phil. Trans. R. Soc. Lond. B* **360**, 815–836 (2005).
28. Lee, T.S. & Mumford, D. Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am A Opt Image Sci Vis.* **20**, 1434–1448 (2003).
29. Kakade, S. & Dayan, P. Acquisition and extinction in autoshaping. *Psychol. Rev.* **109**, 533–544 (2002).
30. Stalnaker, T.A., Berg, B., Aujla, N. & Schoenbaum, G. Cholinergic interneurons use orbitofrontal input to track beliefs about current state. *J. Neurosci.* **36**, 6242–6257 (2016).
31. Takahashi, Y.K., Langdon, A.J., Niv, Y. & Schoenbaum, G. Temporal specificity of reward-prediction errors signaled by putative dopamine neurons in rat VTA depends on ventral striatum. *Neuron* **91**, 182–193 (2016).
32. Ludvig, E.A., Sutton, R.S. & Kehoe, E.J. Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Comput.* **20**, 3034–3054 (2008).
33. Gershman, S.J., Moustafa, A.A. & Ludvig, E.A. Time representation in reinforcement-learning models of the basal ganglia. *Front. Comput. Neurosci.* **7**, 194 (2014).
34. Mello, G.B.M., Soares, S. & Paton, J.J. A scalable population code for time in the striatum. *Curr. Biol.* **25**, 1113–1122 (2015).

ONLINE METHODS

Mice. We used seven adult male mice that were heterozygous for the gene expressing the Cre recombinase under the control of the promoter from the solute carrier family 6 (neurotransmitter transporter, dopamine), member 3 gene (*Slc6a3*) (B6.SJL-Slc6a3$^{\mathrm{tm1.1(cre)Bkmm}}$/J mice; The Jackson Laboratory) and backcrossed for >5 generations with C57/BL6J mice[35]. Three mice were used in task 1 (**Fig. 1a**), one mouse was used in task 2 (**Fig. 1b**), and three mice were used in task 2b (**Supplementary Fig. 1**). The mice were housed on a 12-h dark:12-h light cycle (dark from 7 a.m. to 7 p.m.). We trained mice on the behavioral task at approximately the same time each day. All experiments were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals[36] and were approved by the Harvard Institutional Animal Care and Use Committee.

Surgery and virus injections. We performed all surgeries under aseptic conditions with mice that were under isoflurane (1–2% at 0.5–1.0 liter/min) anesthesia. Analgesic (buprenorphine, 0.1 mg per kg body weight; by intraperitoneal injection) was administered pre-operatively and at 12-h checkpoints post-operatively. We performed two surgeries, both stereotactically targeting the left VTA (from bregma: 3.1 mm posterior, 0.6 mm lateral, 4.2 mm ventral). In the first surgery, we injected 500 nl of adeno-associated virus serotype 5 (AAV5) carrying an inverted ChR2-encoding sequence (H134R) fused to the sequence expressing the fluorescent reporter eYFP and flanked by double *loxP* sites[3,37]. We previously showed that the expression of this virus is highly selective and efficient in dopamine neurons[3]. After 2 weeks, we performed the second surgery to implant a head plate and custom-built microdrive containing 6–8 tetrodes and an optical fiber.

Behavioral paradigm. After 1 week of post-surgical recovery, we water-restricted mice in their cages. Weight was maintained above 85% of the body weight prior to the water restriction. We habituated and briefly head-restrained mice for 2–3 d before training. Odors were delivered to mice with a custom-made olfactometer[38]. Each odor-emitting compound was dissolved in mineral oil at 1/10 dilution. 30 μl of diluted odor-emitting compound was placed on a glass fiber filter paper. Filtered air was run through the filter paper to produce a total flow rate of 1 liter/min. Odor-emitting compounds included isoamyl acetate, (+)-carvone, 1-hexanol, *p*-cymene, ethyl butyrate, 1-butanol, limonene, dimethoxybenzene, caproic acid, 4-heptanone and eugenol. The combination of these odors differed for different mice. We automatically detected licks by measuring breaks in an infrared beam that was placed in front of the water spout.

For both tasks, rewarded odor A trials consisted of 1s odor presentation followed by a delay chosen from a Gaussian distribution defined over nine points (1.2 s 1.4 s, 1.6 s, 1.8 s, 2.0 s, 2.2 s, 2.4 s, 2.6 s and 2.8 s; mean = 2 s; s.d. = 0.5 s) prior to reward delivery. For both tasks 1 and 2, rewarded odor B and odor C trials consisted of a 1-s odor presentation followed by either a 1.2-s or 2.8-s delay from odor onset, respectively, prior to reward delivery (**Fig. 1a**,**b**). In task 2b, rewarded odor B trials consisted of a 1-s odor presentation followed by a 2-s delay from odor onset; odor C was not given (**Supplementary Fig. 1**). In all tasks, odor D trials were unrewarded. In task 1, reward was given in 100% of trials. In tasks 2 and 2b, reward was given in 90% of trials. For all of the tasks, reward size was kept constant at 3 μl. Trial type was drawn pseudo-randomly from a scrambled array of trial types, to keep the proportion of trial types constant between sessions. The ITI between trials was drawn from an exponential distribution (mean = 12–14 s) to ensure a flat hazard function. Mice performed between 150 and 300 trials per session.

Electrophysiology. We based recording techniques on previous studies[3,4,13]. We recorded extracellularly from the VTA using a custom-built, screw-driven Microdrive (Sandvik, Palm Coast, Florida) containing eight tetrodes glued to a 200-μm optic fiber (ThorLabs). Tetrodes were glued to the fiber and clipped so that their tips extended 200–500 μm from the end of the fiber. We recorded neural signals with a DigiLynx recording system (Neuralynx) and data acquisition device (PCIe-6351, National Instruments). Broadband signals from each wire were filtered between 0.1 and 9,000 Hz and recorded continuously at 32 kHz. To extract spike timing, signals were band-pass-filtered between 300 and 6,000 Hz and sorted offline using MClust-3.5 (A.D. Redish). At the end of each session, the fiber and tetrodes were lowered by 75 μm to record new units the next day. To be included in the data set, a neuron had to be well isolated (L-ratio (a measure of unit isolation quality) < 0.05)[39] and recorded within 300 μm of a light-identified dopamine neuron (see next paragraph) to ensure that it was recorded in the VTA. We also histologically verified recording sites by creating electrolytic lesions using 10–15 s of 30-μA direct current.

To unambiguously identify dopamine neurons, we used ChR2 to observe laser-triggered spikes[3,40,41]. The optical fiber was coupled with a diode-pumped solid-state laser with analog amplitude modulation (Laserglow Technologies). At the beginning and end of each recording session, we delivered trains of ten 473-nm light pulses, each 5 ms long, at 1, 5, 10, 20 and 50 Hz, with an intensity of 5–20 mW/mm^2 at the tip of the fiber. Spike shape was measured using a broadband signal (0.1–9,000.0 Hz) sampled at 32 kHz. To be included in our data set, neurons had to fulfill three criteria[3,4,13]. (i) The neurons' spike timing must be significantly modulated by light pulses. We tested this by using the stimulus-associated spike latency test (SALT)[41]. We used a significance value of $P < 0.05$, and a time window of 10 ms after laser onset. (ii) Laser-evoked spikes must be near-identical to spontaneous spikes. This ensured that light-evoked spikes reflected actual spikes instead of photochemical artifacts. All light-identified dopamine neurons had correlation coefficients > 0.9 (**Supplementary Fig. 3b**,**g**). (iii) Neurons must have a short latency to spike following laser pulses, and little jitter in spike latency (**Supplementary Fig. 3c**,**e**,**f**). Although others have used a latency criteria of 5 ms or less ('short latency')[3,4,13], we found that the high laser intensity required to elicit this short latency spike sometimes created a mismatched waveform, due to two neurons near the same tetrode being simultaneously activated. For this reason, we often decreased the laser intensity and elicited a spike 5–10 ms ('longer latency') after laser onset. We separately analyzed neurons in both the short-latency and longer-latency categories, and we found qualitatively similar results in each group. Therefore, we pooled all dopamine neurons with latencies below 10 ms in our analyses.

Data analysis. We focused our analysis on light-identified dopamine neurons (n = 30 for task 1; n = 43 for task 2). To measure firing rates, PSTHs were constructed using 1-ms bins. Averaged PSTH values shown in figures were smoothed with a box filter of 100–150 ms. Average pre-reward firing rates were calculated by counting the number of spikes 0–400 ms prior to reward onset. We also attempted to use window sizes ranging from 200 to 500 ms, and these produced similar results. Average post-reward firing rates were calculated by counting the number of spikes 50–300 ms after reward onset in both tasks 1 and 2. Both pre-reward and post-reward responses were baseline-subtracted, with a baseline reading taken 0–1 s prior to odor onset.

We further examined the licking behavior on each day of recording. We fit a logistic function to each day's data, for each animal, which took the following form:

$$f(t)=\frac{L}{1+\mathrm{e}^{-k(t-t_0)}}$$

where t is time relative to odor onset, L is the curve's maximum value, k is the steepness of the curve and t_0 is the time of the sigmoid's midpoint.

We plotted a subjective hazard rate by blurring the probability distribution function $p(t)$ by a normal distribution whose s.d. scales with elapsed time. Similar to previous work[23,24], we used a Weber fraction $\phi = 0.25$:

$$\tilde{p}(t)=\frac{1}{\phi t\sqrt{2\pi}}\int_{-\infty}^{\infty} f(\tau)e^{-(\tau-t)^2/(2\phi^2 t^2)}d\tau$$

Statistical analyses. No statistical methods were used to pre-determine sample sizes, but our sample sizes were similar to those reported in previous publications[3,4,13]. Data collection and analysis were not performed in a manner blinded to the conditions of the experiments. Mice were chosen at random for task 1 or task 2. All trial types were randomly interleaved within a single recording session. We verified that all groups of data (including both electrophysiology and behavior) that were compared using ANOVAs did not deviate significantly from a normal distribution, using a chi-square goodness-of-fit test. To test whether dopamine RPEs were modulated by ISI length, we used a two-factor ANOVA, with a neuron and ISI as factors. To test whether licking was modulated by odor identity, we used a one-factor ANOVA, with odor identity as a factor. To test whether individual neurons' RPEs were modulated by factors such as ISI length

doi:10.1038/nn.4520

or lick rates, we fit a line to the data (dopamine RPEs versus ISI) and reported the slope. We also displayed the 95% CI of the slope (**Supplementary Fig. 3**) or a summary of whether or not the 95% CI included 0 (shaded in **Fig. 3g**,**h**). A **Supplementary Methods Checklist** is available.

Code availability. Code used to implement the computational modeling in this manuscript can be found at https://github.com/cstarkweather.

Immunohistochemistry. After 4–8 weeks of recording, we injected mice with an overdose of ketamine and medetomidine. Mice were exsanguinated with saline and perfused with 4% paraformaldehyde. We cut brains in 100-μm coronal sections on a vibrotome and immunostained them with antibodies to tyrosine hydroxylase (AB152, 1:1,000; Millipore) to visualize dopamine neurons. We additionally stained brain slices with 49,6-diamidino-2-phenylindole (DAPI; Vectashield) to visualize nuclei. We confirmed AAV expression with eYFP fluorescence. We examined slides to verify that the optic fiber track and electrolytic lesions were located in a region with VTA dopamine neurons and in a region expressing the AAV (**Supplementary Fig. 7**).

Computational modeling. *Temporal difference (TD) model.* We first simulated TD error signaling in our tasks by using TD learning with a CSC representation, identical to the algorithm presented by Schultz and colleagues[1]. We set stimulus onset at $t = 20$ arbitrary units of time and set nine possible reward times at $t = 26, 27, 28, 29, 30, 31, 32, 33$ and 34 arbitrary units of time. Each arbitrary unit of time in this model corresponded to 200 ms in our task. In our task 1 simulation, reward was always delivered. In our task 2 simulation, reward was delivered in 90% of trials. The results presented in the text were obtained by running 10× simulations of each task, with 5,000 trials per simulation. The 'TD-with-reset' variant was simulated by setting the error term to 0 at any of the time points after reward was delivered.

Belief-state TD model. We next simulated TD error signaling in our tasks by using a belief-state TD model, similar to that proposed by Daw and colleagues, as well as by Rao[8,9]. To capture the discrete dwell times in our tasks (1-s odor presentation, followed by nine discrete possible reward delivery timings at 1.2, 1.4, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6 and 2.8 s after odor onset), we coded a Markov equivalent of a semi-Markov model[9]. This Markov equivalent contained 15 total hidden sub-states (**Supplementary Fig. 7a**,**b**). Sub-states 1–5 corresponded to the passage of time during the 1-s odor presentation; sub-states 6–14 corresponded to the passage of time preceding the nine possible reward delivery time. Sub-state 15 corresponded to the ITI. If reward was received at the earliest possible time (1.2 s), then this would correspond to the model proceeding through sub-states 1–6, and then transitioning to sub-state 15. If reward was received at the latest possible time (2.8 s), then this would correspond to the model proceeding through sub-states 1–14, and then transitioning to sub-state 15.

In the belief-state TD model, it is assumed that the animal has learned a state transition distribution, encoded by matrix T. We captured the dwell-time distribution in the ISI state by setting elements of T to match either the hazard function or the inverse hazard function of receiving reward at any of the nine discrete time points. For example, the hazard rate of receiving reward at 1.2 s would correspond to $T(6,15)$, or the probability of transitioning from sub-state 6→15. 1 minus the hazard rate of receiving a reward at 1.2 s would correspond to $T(6,7)$, or the probability of transitioning from sub-state 6→7. We captured the exponential distribution of dwell times in the ITI state by setting $T(15,15) = 64/65$, and $T(15,1) = 1/65$. An exponential distribution with a hazard rate ('ITI_hazard') of 1/65 has an average dwell time of 65 arbitrary units of time. This average ITI dwell time was proportionally matched to the average ISI dwell time to be comparable to our task parameters. The only difference in T between task 1 and task 2 was as follows:

Task 1:
$T(15,15) = 1 - \text{ITI\_hazard}$
$T(15,1) = \text{ITI\_hazard}$

Task 2:
$T(15,15) = 1 - (\text{ITI\_hazard} \times 0.9)$
$T(15,1) = \text{ITI\_hazard} \times 0.9$

This difference in T between tasks 1 and 2 captured the probability of undergoing a hidden-state transition from ITI back to the ITI, in the case of 10% omission trials. In the belief-state TD model, it is also assumed that the animal has learned a probability distribution over observations given the current state, encoded by observation matrix O. There were three possible observations: null, cue and reward. The likelihood of a particular observation given that the hidden state underwent a transition from $i \rightarrow j$ was captured as follows:

$O(i,j,1)$ = likelihood of observation of 'null', given $i \rightarrow j$ transition
$O(i,j,2)$ = likelihood of observation of 'cue', given $i \rightarrow j$ transition
$O(i,j,3)$ = likelihood of observation of 'reward', given $i \rightarrow j$ transition

To switch from sub-state 15 (ITI) to sub-state 1 (first state of ISI), the animal must have an observation of the cue: $O(15,1,2) = 1$. To switch from sub-state 10 (middle of ISI) to sub-state 15 (ITI), the animal must have an observation of reward: $O(10,15,3) = 1$. The only difference in O between task 1 and task 2 was as follows:

Task 1:
$O(15,15,1) = 1$ (null observation)

Task 2:
$O(15,15,1) = 1 - (\text{ITI\_hazard} \times 0.1)$ (null observation)
$O(15,15,2) = \text{ITI\_hazard} \times 0.1$ (cue in a small percentage of cases)

This difference in O between tasks 1 and 2 captures the fact that in 10% omission trials the animal will observe a cue, but in fact be in the hidden-ITI state rather than a hidden-ISI state.

The results presented in the text were produced by training the belief-state TD model on either task 1 (100% rewarded) or task 2 (90% rewarded), for 5,000 trials each. We found that the model yielded asymptotic results after about 1,000 trials. For this reason, the results shown in the text are taken from trials 2,000–5,000. In all simulations, we used a learning rate of $\alpha = 0.1$ and a discount factor of $\gamma = 0.98$.

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

35. Backman, C. *et al.* Characterization of a mouse strain expressing Cre recombinase from the 3′ untranslated region of the dopamine transporter locus. *Genesis* **44**, 383–390 (2006).
36. National Research Council. *Guide for the Care and Use of Laboratory Animals* 8th edn. (The National Academies Press, 2011).
37. Atasoy, D., Aponte, Y., Su, H.H. & Sternson, S.M. A FLEX switch targets channelrhodopsin-2 to multiple cell types for imaging and long-range circuit mapping. *J. Neurosci.* **28**, 7025–7030 (2008).
38. Uchida, N. & Mainen, Z.F. Speed and accuracy of olfactory discrimination in the rat. *Nat. Neurosci.* **6**, 1224–1229 (2003).
39. Schmitzer-Torbert, N., Jackson, J., Henze, D., Harris, K. & Redish, A.D. Quantitative measures of cluster quality for use in extracellular recordings. *Neuroscience* **131**, 1–11 (2005).
40. Lima, S.Q., Hromádka, T., Znamenskiy, P. & Zador, A.M. PINP: a new method of tagging neuronal populations for identification during *in vivo* electrophysiological recording. *PLoS One* **4**, e6099 (2009).
41. Kvitsiani, D. *et al.* Distinct behavioral and network correlates of two interneuron types in prefrontal cortex. *Nature* **498**, 363–366 (2013).