Hidden State Inference in the Midbrain Dopamine System

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th>Citable link</th>
<th><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:40049995">http://nrs.harvard.edu/urn-3:HUL.InstRepos:40049995</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Hidden State Inference in the Midbrain Dopamine System

A dissertation presented

by

Clara Starkweather

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Neurobiology

Harvard University

Cambridge, Massachusetts

March 2018
Hidden State Inference in the Midbrain Dopamine System

Abstract

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’).

In Chapter 1, I asked whether dopaminergic signaling supports a TD learning framework that operates over hidden states. I found that dopamine signaling exhibited a striking difference between two tasks that differed only with respect to whether reward was delivered deterministically. My results favor an associative learning rule that combines cached values with hidden state inference. In Chapter 2, I used the behavioral paradigm developed in Chapter 1 to examine a possible cortical contribution to computing dopamine RPEs. I found that inactivation of the medial prefrontal cortex (mPFC) affected dopaminergic signaling in a task in which the state of the environment was hidden and must be inferred, but not in a task in which the state was known with certainty. Computational modeling suggests that the effects of inactivation are best explained by a circuit in which the mPFC conveys inference over hidden states to the dopamine system.
My findings support a reinforcement learning circuitry in which the mPFC furnishes inferences about hidden states into the subcortical reward prediction machinery, with dopamine neurons signaling errors in these reward predictions. These results provide key insights into how the brain implements reinforcement learning, particularly in ambiguous settings.
# Table of Contents

Introduction...........................................................................................................................................1

Chapter 1. Dopamine Reward Prediction Errors Reflect Hidden State Inference Across Time.........................................................................................................................21

Chapter 2. Medial Prefrontal Cortex Shapes Dopamine Reward Prediction Errors Under State Uncertainty........................................................................................................58

Conclusion................................................................................................................................................110

References................................................................................................................................................120
# Table of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>Dopamine reward prediction error signals and TD learning</td>
<td>8</td>
</tr>
<tr>
<td>0.2</td>
<td>CSC versus microstimulus features</td>
<td>11</td>
</tr>
<tr>
<td>0.3</td>
<td>Belief state TD model</td>
<td>15</td>
</tr>
<tr>
<td>1.1</td>
<td>Task design</td>
<td>24</td>
</tr>
<tr>
<td>1.2</td>
<td>Mice learn to lick in anticipation of reward</td>
<td>25</td>
</tr>
<tr>
<td>1.3</td>
<td>Recording sites</td>
<td>26</td>
</tr>
<tr>
<td>1.4</td>
<td>Optogenetic identification of dopamine neurons</td>
<td>26</td>
</tr>
<tr>
<td>1.5</td>
<td>Averaged dopamine activity in Tasks 1 and 2 shows different patterns of modulation over variable ISI interval</td>
<td>27</td>
</tr>
<tr>
<td>1.6</td>
<td>Onset of phasic post-reward RPEs begins 50ms following water valve opening</td>
<td>28</td>
</tr>
<tr>
<td>1.7</td>
<td>Individual dopamine neurons show opposing patterns of post-reward firing in Tasks 1 and 2</td>
<td>29</td>
</tr>
<tr>
<td>1.8</td>
<td>TD with CSC model, with or without reset, is inconsistent with our data</td>
<td>33</td>
</tr>
<tr>
<td>1.9</td>
<td>Belief state model parameters</td>
<td>34</td>
</tr>
<tr>
<td>1.10</td>
<td>Belief state model is consistent with our data</td>
<td>36</td>
</tr>
<tr>
<td>1.11</td>
<td>Belief state model shapes value signals that different between Tasks 1 and 2, leading to opposite patterns of post-reward modulation over time</td>
<td>37</td>
</tr>
<tr>
<td>1.12</td>
<td>Belief state model reproduces modulation of pre- and post-reward RPEs over variable ISI interval</td>
<td>38</td>
</tr>
<tr>
<td>1.13</td>
<td>Post-reward RPEs normalized to free water response</td>
<td>39</td>
</tr>
<tr>
<td>1.14</td>
<td>Task 2b</td>
<td>40</td>
</tr>
<tr>
<td>1.15</td>
<td>Belief state model result is agnostic to precise shape of trained distribution</td>
<td>41</td>
</tr>
</tbody>
</table>
Table 2.1 Summary of animals used in Chapter 2 97

Figure 3.1 Recurrent network implementation of a belief state 115

Figure 3.2 Compressing belief state dimensions 117
Acknowledgements

I would first like to thank my wonderful PI, Nao Uchida. Nao is a pleasure to discuss science with. During my PhD, I was constantly inspired by Nao’s enthusiasm for uncovering the ‘big picture’ implications of our work. Without his encouragement, I would not have delved into models of reinforcement learning as deeply, and would have missed out on the valuable (and unexpected) computational neuroscience aspects of my PhD. Yet, at the same time, Nao is a practical and grounded mentor. Any time an experiment was not working, even when I was quite discouraged, I would always leave Nao’s office uplifted and armed with practical guidance. I think it’s rare to find a mentor with both of these qualities. I am grateful and honored to have spent the past few years working with Nao Uchida.

Next, I want to thank Sam Gershman, who has been a great collaborator. Sam patiently helped me to implement the reinforcement learning models I discuss in my dissertation. My understanding of my data’s implications for reinforcement learning theory has been greatly elevated by our collaboration.

I also want to thank Mitsuko Watabe-Uchida, who has helped tremendously on technical aspects of my project. For instance, Mitsuko encouraged me to try the second-generation chemogenetic approach that I discuss in my dissertation, and helped me find channels of communication for making these experiments work.

My committee members—Rick Born, John Assad, and Chris Harvey—have provided valuable feedback throughout my PhD. My discussions with my committee were some of the highlights of my time in graduate school. These discussions encouraged me to consider other ways of modeling my data, and forced me to clearly rationalize my choices of particular interpretations.

Finally, I would not have been able to complete the experiments or analyses in my dissertation without the technical guidance of my labmates. In particular, I thank Mackenzie Mathis, Will Menegas, Hideyuki Matsumoto, and Ju Tian, for surgery and drive-building trainings and tips. I thank Hideyuki Matsumoto for showing me how to identify dopamine neurons using optogenetics. I thank Ju Tian for sharing her strategy for efficiently storing single-unit recording data. I thank all of my labmates for providing feedback during lab meetings, and for making our lab office a pleasure to work in. I will miss working in such a joyful and kind atmosphere!
INTRODUCTION
INTRODUCTION

In order to produce adaptive behaviors, animals must learn to predict relevant outcomes in their environments. Certain outcomes (‘unconditioned stimuli’) trigger a biological response in the absence of learning, such as salivation accompanying a bite of tasty food. During classical conditioning, animals learn to predict an unconditioned stimulus, or US, following a neutral sensory stimulus that does not evoke an automatic biological response (‘conditioned stimulus’, or CS). This type of prediction is critical in order to optimize choices and prepare behavioral reactions. In this introduction, I will outline computational models of associative learning. I will focus mostly on classical conditioning tasks, which have proven to be a valuable paradigm for studying associative learning models and their neural correlates. I will explain how recent neurophysiology data have put various models to the test. These data have led to proposed changes in these models, as well as new questions to be addressed through experiments. My dissertation work will directly test some of these emergent questions.

Pioneering theories of reinforcement learning

Two broad categories of behavioral paradigms spurred theories of associative learning. One category was classical conditioning, in which animals are presented with CS-US pairings. The US is delivered regardless of the animal’s behavior during the time between the two stimuli. In the famous example of Pavlov’s dog, a bell was rung (CS) every time the animal was presented with meat powder (US) (Pavlov, 1927). After repeating this procedure multiple times, the CS alone was sufficient to elicit conditioned responding (salivation), even in the absence of the US. Another category of behavioral paradigms was operant conditioning, in which animals must generate an action in order to achieve an outcome. During operant conditioning, certain
actions become entrained to cues in the environment, called discriminative stimuli. A classic example of operant conditioning was Thorndike’s puzzle box, from which cats would learn to escape in order to reach a scrap of fish (Thorndike, 1911). Thorndike observed that inexperienced cats would produce a huge variety of actions in order to escape: squeezing, clawing, and striking, among others. Actions that tended to be associated with favorable outcomes (escape and food receipt) were performed at higher rates in subsequent trials. Over time, once cats were faced with the sensory environment of the box, they immediately performed the actions necessary to escape. Based on his observations, Thorndike posited a ‘law of effect’: actions are potentiated if they produce favorable outcomes.

The formation of associations in classical conditioning, however, could not be explained by a stimulus-response association driven by the outcome’s favorability, such as that described by Thorndike, because the outcome would be delivered regardless of the animals’ behavior. One early theory, developed by Bush and Mosteller, provided a mathematical account for strengthening the behavioral association between a CS and US (Bush and Mosteller, 1955). This theory held that the spatiotemporal contiguity of the CS and US was necessary and sufficient for developing a behavioral association. The Bush Mosteller model set the change in the probability of conditioned responding to CS A (ΔpA) directly proportional to the discrepancy between the maximum possible probability of responding to the US (λ) and current probability of responding to CS A (pA):

\[ \Delta p_A = \alpha(\lambda - p_A) \]  

(1)

In Equation 1, \( \alpha \) is a constant with a value between 0 and 1 that reflects the animal’s learning rate. Prior to training, \( p_A \) is 0. Thus, \( (\lambda - p_A) \), is large once the US is received, driving large
changes in \( p_A \) on the first experienced trials. As \( p_A \) increases over trials, \((\lambda - p_A)\) decreases, resulting in smaller incremental changes in \( p_A \) with each subsequent trial. Therefore, the Bush-Mosteller model neatly describes the incremental increase in conditioned responding, eventually reaching an asymptote, that is observed across trials in classical conditioning experiments.

In 1968, Kamin reported a behavioral phenomenon in rats that could not be explained by Bush and Mosteller’s account (Kamin, 1968). First, CS A is repeatedly paired with a reward. Following this phase of training, \( p_A \) should be high. In the next phase of training, CS A and CS B are both repeatedly paired with reward. During this phase of training, the weight update \( \alpha(\lambda - p_B) \) should be positive, driving up the value of \( p_B \). Yet, in the testing phase, it was revealed that presentation of CS A, but not CS B, elicits conditioned responding, inconsistent with the Bush-Mosteller model. However, a theory developed by Rescorla and Wagner could explain this phenomenon (Rescorla and Wagner, 1972). In the Rescorla Wagner model, the change in associative strength between CS A and the US (\( \Delta V_A \)) is directly proportional to the discrepancy between the maximum possible responding the US (\( \lambda \)) and the current associative strength of all the CS’s present (\( V_T \)):

\[
\Delta V_A = \alpha(\lambda - V_T)
\]

In Equation 2, \( \alpha \) is a constant with a value between 0 and 1 that reflects the animal’s learning rate. While the form of this equation is very similar to Equation 1, there are two critical changes. First, the Rescorla Wagner model explicitly refers to associative strengths \( V \), rather than probabilities of response. Second, a cue’s associative strength is updated proportional to the discrepancy between the maximal responding ascribed to the US (\( \lambda \)), and the predicted associative strength based on all of the CS’s present. The first phase of training produces a value
$V_A$ that matches $\lambda$. During the second phase of training, the weight update $\alpha(\lambda - V_T)$ equals zero because $V_T$ includes both CS A (which has a value already equal to $\lambda$) and CS B. No additional associative strength is gained for CS B because CS A already predicts the reward, thereby explaining why CS B elicited no conditioned responding on its own. In this way, the Rescorla Wagner model asserts that no new learning will occur unless an existing prediction is violated. An error between the actual and predicted outcome is needed in order to drive learning.

**Temporal difference learning and dopamine signaling**

The idea that an error between actual and predicted outcome drives learning has since been at the center of associative learning theories. In 1988, a related learning theory was born in the field of computer science (Sutton, 1988). This theory, called Temporal Difference learning (TD learning), also used the discrepancy between actual and expected outcome to drive learning, similar to the Rescorla-Wagner model. However, in a critical departure, TD learning formally defined value as the discounted sum of future reward:

$$V(t) = \sum_{\tau=t}^{\infty} \gamma^{\tau-t} r(\tau)$$

(3)

where $t$ is the current time, $r(\tau)$ is the reward at time $\tau$, and $\gamma$ is a discount factor ($0 \leq \gamma \leq 1$) that down-weights future rewards. This definition of value allows learning to occur at every timepoint, rather than only at the moment when reward is delivered (as was the case in the Rescorla Wagner model), because differences in actual and expected value may occur continuously throughout time. The value function estimate is modeled as a linear combination of stimulus features, which classically corresponds to a ‘Complete Serial Compound (CSC)’ representation that tracks elapsed time relative to observable stimuli:
\[ \hat{V}(t) = \sum_i w_i x_i(t) \] (4)

where \( x(t) \) represents CSC features, and \( w_i \) is a predictive weight associated with feature \( i \). \( x_1(t) \) is a vector of zeros, except for a value of 1 for a single timepoint 1 relative to cue onset. \( x_2(t) \) is an identical vector, except that the value of 1 would occur at timepoint 2. In this way, the CSC features \( x(t) \) show serial activations that span time intervals (Figure 0.1b). The weights are updated according to the following learning rule:

\[ \Delta w_i = \alpha x_i(t) \delta(t) \] (5)

where \( \alpha \) is a learning rate \( (0 \leq \alpha \leq 1) \) and \( \delta(t) \) is the prediction error in the value signal. The discrepancy between actual and predicted value is computed according to:

\[ \delta(t) = r(t) + \gamma \hat{V}(t + 1) - \hat{V}(t) \] (6)

where \( r(t) \) represents reward at time \( t \). In order to gain an intuition for computing TD errors \( \delta(t) \), it is useful to note that the latter two terms of Equation 6—\( \gamma \hat{V}(t + 1) - \hat{V}(t) \)—resemble the temporal derivative of the value estimate, if \( \gamma \) is close to 1. In other words, the temporal difference error \( \delta(t) \) can be approximated by the temporal derivative of the value estimate at time \( t \), plus the reward at time \( t \).

Before learning, the TD model has not learned any weights \( w \). Therefore, the value estimate \( \hat{V}(t) \) is zero at all timepoints. When the model experiences US delivery at time \( t \), this produces a positive prediction error \( \delta(t) \). This positive prediction error increases the weight associated with the corresponding CSC feature \( x_t \), resulting in a positive value estimate at time \( t \). Thus, the temporal derivative of the value estimate becomes positive at time \( t - 1 \), producing a positive prediction error \( \delta(t - 1) \) and increasing the weight associated with the corresponding
CSC feature $x_{t-1}$. Over trials, the positive prediction errors propagate back in time, increasing the weights of CSC features that fully tile the interstimulus interval (‘ISI’), until the only positive prediction error occurs at the time of the CS (if CS onset is unpredictable). In this way, the TD algorithm eventually learns a value estimate that precisely spans the time between the CS and US (Figure 0.1c). For clarity, I will refer to this TD algorithm as the CSC TD algorithm.

Although TD learning was originally developed for computer science applications, neurophysiology experiments in the 1990’s by Schultz and colleagues revealed an exceptional correspondence between the error signals predicted by TD learning and those signaled by midbrain dopamine neurons (Mirenowicz and Schultz, 1994; Hollerman and Schultz, 1996; Schultz et al, 1997; Hollerman and Schultz, 1998; Bayer and Glimcher, 2005; Figure 0.1a,c). Specifically, the authors trained monkeys on a classical conditioning task in which a CS was followed by a US, slightly more than 1s apart from one another. Before learning, dopamine neurons showed phasic activation only at the time of the US. After learning, this phasic dopamine response instead occurred at the time of the CS, and the signal at the time of the US was much smaller than before learning. Strikingly, if the US was unexpectedly omitted, dopamine neurons briefly paused their tonic firing exactly at the time of the usual US delivery. These signals, collectively referred to as Reward Prediction Error (‘RPE’), match the error signals proposed by the CSC TD learning algorithm (Figure 0.1a,c):

(1) Before learning, the only positive TD error $\delta(t)$ is at the time of the US due to positive excitation from reward $r(t)$. 
(2) After learning, the sustained value signal commences following the CS, resulting in a positive TD error $\gamma \hat{V}(t + 1) - \hat{V}(t)$ and therefore a positive TD error at CS onset.

(3) After learning, because the sustained value prediction drops to zero at the time of the predicted US, the TD error $\gamma \hat{V}(t + 1) - \hat{V}(t)$ is negative at the time of the US and cancels out positive excitation from reward $r(t)$, resulting in a smaller response at the time of a predicted reward.

(4) After learning, if the US is not delivered at the predicted time, the negative TD error is not cancelled out by the arrival of reward, resulting in a negative signal.

In these four ways, the firing patterns of dopamine neurons are recapitulated in TD error signals.

It is important to note that the Rescorla Wagner model would not capture all four of these

---

**Figure 0.1** | **Dopamine reward prediction error signals and TD learning.** (a) Firing pattern of a single putative dopaminergic neuron in the midbrain of a primate during a classical conditioning task. From Schultz et al, 1997. (b) Temporal difference (TD) model architecture with complete serial compound (CSC) feature representation. (c) Value signal, temporal difference of value signal, and error signal produced by TD model with the CSC.
characteristics. The Rescorla Wagner model would predict the US because it is associated with the CS, resulting in a smaller error signal at the time of the US. However, the Rescorla Wagner model does not have a timing mechanism, and therefore would not capture the negative signal observed on reward omission at the usual time of the US. The correspondence between TD errors and dopamine signals led to the hypothesis that TD learning approximates how reinforcement learning is actually implemented in the brain. In line with this hypothesis, other theoretical works speculated that the cortex conveys the CSC temporal features into the striatum, where value is computed (Houk et al, 1995). By modulating the weights of corticostriatal synapses, dopamine signals could plausibly shape the striatal value representation. In this way, the TD learning model provided a cornerstone for understanding how reinforcement learning could be biologically implemented.

**Timing in temporal difference (TD) models**

The CSC TD model captures the exquisite temporal specificity of dopamine reward prediction errors. First, dopamine neurons show a negative prediction error exactly at the time of an expected reward. Second, dopamine responses to reward are suppressed only at the time of the expected reward. In a 1998 study, Hollerman and Schultz occasionally shifted the timing of reward by just 500ms earlier or later than the time of the usual reward delivery. This small temporal jitter was sufficient to evoke a larger dopamine response than if reward were delivered at its usual time. This finding is consistent with the errors produced by the TD model, because the temporal difference of the value signal is most negative at the cue-reward delay time that the animal was trained on (thereby canceling excitation from reward only at that time, see Figure 0.1c). Finally, another timing-related characteristic of dopamine RPEs, which is captured by the TD model, is delay discounting. Because value corresponds to the discounted sum of future
rewards (Equation 3), the magnitude of the value signal at cue onset is inversely related to the delay between cue and reward. Consistent with this, several studies in both rodents and primates have observed delay discounting in the cue response of dopamine neurons. In these studies, different cues predicted different CS-US delay times (Kobayashi and Schultz, 2008; Fiorillo et al, 2008; Starkweather et al, 2017). Cues followed by late rewards result in smaller dopamine responses than cues followed by early rewards. In this way, the CSC TD framework matches several findings relating to timing and the dopamine system.

However, additional experimental observations suggest that the CSC TD model does not provide a complete account of timing in the dopamine system. The first observation is that dopamine responses are not thoroughly suppressed at the time of expected rewards, particularly for CS-US pairings involving long ISIs. This is the case even in primates trained extensively on temporally precise visual cues. In two 2008 studies, the authors trained animals of delay times ranging from 1s to 16s (Kobayashi and Schultz, 2008; Fiorillo et al, 2008). US responses were smallest (most suppressed) for 1s ISI trial types, and largest (least suppressed) for 16s ISI trial types. A CSC TD model would suppress the US responses equally for all trials, irrespective of the delay time. The second observation is that the ‘dip’ observed upon reward omission is temporally extended (for up to 1 second; Schultz et al, 1997; Matsumoto and Hikosaka, 2007; Tian and Uchida, 2015). In contrast, a CSC TD model produces a sharp dip only at the time of the expected reward (Figure 0.1c). Based on these experimental observations, theorists have proposed a TD model that includes temporally blurred ‘microstimulus’ features (Ludvig et al, 2008).
The microstimulus TD model swaps the CSC representation for Gaussian distributions, whose widths increase over elapsed time (Figure 0.2). The areas under the various Gaussian distributions are equal, so the increasing widths mean that the heights of the distributions must decrease over elapsed time. The increasing widths of the microstimuli respect Weber’s law as it applies to interval timing. This law stipulates that the variance of timing estimation increases linearly with elapsed time, and agrees well with behavioral studies of interval timing (Balsam and Gallistel, 2009). In other words, as timing estimation becomes noisier over elapsed time, the widths of microstimuli increase. If a reward occurs at a long delay from CS onset, the reward increases the weights of more than one microstimulus feature (contrasting with the CSC TD model), because the Gaussian distribution of multiple features will have nonzero values at the moment that a reward is received. Accordingly, the TD model increases the weights of multiple microstimulus features that are active at that time. Over many trials, this produces a value signal that is less sharply resolved in time than the CSC TD model (Figure 0.2). Specifically, the value function will be positive at more timepoints than just the exact moment that reward is received, and it will have a smaller amplitude (due to the decreasing heights of the microstimuli, particularly for longer reward delays). Therefore, the TD error, $\gamma \hat{V}(t + 1) - \hat{V}(t)$, will be both shallower and more spread out over time, and this shallowness and temporal spread become more pronounced for

![Figure 0.2 | CSC versus microstimulus features.](image-url)

The width of microstimulus features expands for longer delay times, resulting in a value signal that is less sharply resolved at later timepoints.
longer delays to reward. As a consequence, when a reward is given, the response cannot be completely canceled out, similar to the data. The reward response is greater for longer delays because the model’s ability to cancel out excitation from the reward becomes less precise for longer intervals. Furthermore, if a predicted reward is omitted, the ‘dip’ is more spread out across time than predicted by the TD model with the CSC. Finally, a last piece of data that can be explained by the microstimulus TD model, is that the baseline tonic firing rate of dopamine neurons decreases slightly before a reward is received. This is a common experimental observation in our lab (Tian and Uchida, 2015; Tian et al, 2016; Starkweather et al, 2017). This pre-reward decrease in firing is reproduced by the microstimulus model because the temporal imprecision in the model results in a negative prediction error commencing prior to the exact time of the reward. Therefore, the microstimulus TD model explains common observations of dopamine recordings, which were inconsistent with the CSC.

Other updates to the TD model have been proposed, which I will discuss in the next section. However, it is important to keep in mind that all these models possess a timekeeping mechanism. Based on the above data and modeling results, dopamine RPEs are likely computed on an imperfect and noisy timing mechanism. Therefore, any modeling refinements proposed in future work should also contain a timing representation that is constrained by scalar timing uncertainty.

State inference in the temporal difference (TD) model

In 1998, Hollerman and Schultz made another important experimental observation in monkeys trained on a CS-US pairing, separated by a constant delay time. In a small proportion of
probe trials, rewards were given earlier than predicted, as discussed above. As would be predicted by the TD model, this early reward produced a large dopamine response because the timing of reward was unexpected. However, the authors made an interesting observation after the early reward had been delivered: there was no omission response at the time of the usual reward. In contrast, the CSC TD model would produce an omission response at the time of the usual reward, regardless of whether an early reward had been delivered (Figure 0.3a,c). Based on this experimental observation, and its discordance with the existing theory, two main categories of modifications were proposed to the TD model.

The first proposed modification was that the TD model simply ‘resets’ the CSC feature representation after a reward is received. There are two possible ways of resetting the feature representation, without altering the CSC. Future errors could be fixed at zero after a reward is received (Suri and Schultz, 1998; Suri and Schultz, 1999). Alternatively, a reward receipt could terminate the set of CSC features that became active during the ISI, and activate a new set of CSC features that became active during the intertrial interval, or ITI (Brown et al, 1999). Implicit in this reset model is that the animal knows (based on reward receipt) that it should no longer expect a reward. This means that the animal infers the ‘state’ of the task based on observable stimuli (the reward), and thus can be considered as a simple example of state inference. If one imagines the task is divided into two states—the ISI state during which a reward is expected, and the ITI state during which a reward is not expected—the receipt of reward initiates CSC features that correspond to the ITI state, and terminates those that correspond to the ISI state. However, this is an ad hoc modification to the TD model with the CSC. It adds a simple form of state inference to the TD model in order to better match the data, but does not acknowledge other
possible instances in which state inference may come into play. For example, what would happen if reward were occasionally omitted? Would the ISI features simply continue on indefinitely because no reward is received? Or would the animal infer over time that a reward is not coming, thereby terminating the ISI features? These issues remain unresolved with the reset model.

The second proposed modification is that the TD model’s features themselves represent the inferred state of the environment (Daw et al, 2006; Rao, 2010). I will refer to this as a belief state TD model. The purely temporal CSC or microstimulus representation is swapped with a belief state, which is an inferred probability distribution over states. Value is no longer proportional to the weight assigned to a particular active state. Rather, value is equal to the weight assigned to a particular state, multiplied by the probability (from the belief state) allotted to that particular state, summed over all possible states:

$$\hat{V}(t) = \sum_i w_i b_i(t)$$  \hspace{1cm} (7)

where $b_i(t)$ represents the belief state (i.e. the probability of being in each state $i$) at time $t$, with $i$ indexing individual states, and $w_i$ is a predictive weight associated with state $i$. Furthermore, weights are updated proportional to the probability with which the agent ‘believes’ it occupies a particular state:

$$\Delta w_i = \alpha b_i(t) \delta(t)$$  \hspace{1cm} (8)
The belief state TD model was originally conceived as a Semi-Markov process involving just two states: the ISI during which reward is expected, and the ITI during which no reward is expected (Figure 0.3b). Semi-Markov dynamics mean that the length of time spent in a particular state is probabilistic and is defined by a probability distribution called a dwell time distribution (as opposed to in a Markov process, where each ‘state’ represents one arbitrary unit of time). For that reason, rewards were discounted by however much time had elapsed in a particular state once they were received, because this time interval could vary from task to task: prediction errors would be larger, therefore driving up the value of a particular state more, if reward arrived very early, whereas the opposite would be true if reward arrived very late.

Figure 0.3 | Belief state TD model (a) The timing of US delivery following CS onset was jittered by 500ms on probe trials. Note that there is no reward omission response if US is given early. From Hollerman and Schultz, 1998. (b) Semi-Markov schematic for computing belief state under fully observable task conditions. Adapted from Daw et al, 2006. (c) Schematic for values and error signals produced by delivering early reward, in belief state TD model vs. CSC TD model. CSC TD model produces a spurious reward omission response after an early reward delivery, which is not observed in the data shown in (a). (d) Semi-Markov schematic for computing belief state under partially observable task conditions. Adapted from Daw et al, 2006. (e) Belief state TD model architecture.
The belief state TD model accounts for the Hollerman and Schultz result, and captures other experimental findings as well. The Hollerman and Schultz experiment was rewarded in 100% of trials, meaning that the task is fully observable. In other words, the state of the task (ISI or ITI) can be reliably deciphered based on sensory cues alone. Upon observing the CS, the belief state allots 100% probability into the ISI state, and 0% probability into the ITI state. Upon observing the US, the belief state allots 100% probability into the ITI state, and 0% probability into the ISI state (Figure 0.3e). If the ISI has accrued a larger weight (as it should, because rewards are only received when the belief state favors the ISI), a transition into the ITI upon receiving the US would immediately result in no future reward expectation. This would abolish the spurious reward omission response reported by the CSC TD model, upon receiving an early reward (Figure 0.3c). A second set of experimental results, which is compatible with the belief state TD model, consists of dopamine responses recorded in 100%-rewarded task contingencies with variable delay times. In these experiments, on any given trial, the delay between CS and US (or a reward-predicting CS) was drawn from a uniform distribution (Fiorillo et al, 2008; Nomoto et al, 2010; Pasquereau and Turner, 2015). Experiments consistently found that rewards delivered earliest in the variable delay interval produced the largest dopamine responses, and rewards delivered later in the interval produced smaller dopamine responses. While the authors argued that these results are explained by expectancy over time resembling a hazard function (the momentary likelihood that an event will occur, given that it hasn’t occurred yet), this result is also compatible with the belief state TD model. Because late rewards within a particular state are discounted more heavily in a Semi-Markov framework, the belief state TD model also captures this pattern of prediction errors across time. In contrast, a CSC TD model, which lacks an explicit representation of state space (ISI vs. ITI), would produce the same prediction error at
each possible time of reward delivery, because the uniform distribution of timings teaches the model to apply the same value prediction at each timepoint.

In these ways, the belief state TD model captures experimental findings from tasks that jitter the timing of reward relative to cue. However, these tasks do not implicate state uncertainty, which is a core tenant of the belief state TD model. Although reward timing was unpredictable in the tasks described above, the state of the task was known with certainty because reward was always delivered. One can imagine the same task utilizing a variable delay time, but a subtle switch of the task contingency—the task being 90%-rewarded, rather than 100%-rewarded. This switch renders the true state of the task hidden (or ‘partially observable’), because observable stimuli (e.g. CS) no longer indicate with 100% probability that the agent is in the ISI state (Figure 0.3d). Upon observing the CS, the belief state allots 90% probability to the ISI state, and 10% probability to the ITI (Figure 0.3e). As time elapses and no reward is received, the inferred probability (computed using Bayes’ theorem) leaks from the ISI state into the ITI state, indicating mounting belief that the present trial is unrewarded. Thus, the belief state evolves differently over time in fully observable (100% rewarded) versus partially observable (<100% rewarded) task conditions, with the latter scenario producing a belief state that becomes increasingly pessimistic over time. Correspondingly, the value function also decreases over time in the 90%-rewarded condition, resulting in TD errors that increase as a function of time.

In Chapter 1, I will directly compare dopamine signals in two task conditions utilizing variable delays, with one task fully observable (100% rewarded) and the other task partially observable (90% rewarded). I will test my empirical results against various model predictions.
The results of this Chapter provide evidence that the reinforcement learning circuitry accesses inferences about the hidden state of the environment, consistent with the belief state TD model.

It is important to note that there are many other types of reinforcement learning tasks in which a belief state could be at play. The example I described above dealt with uncertainty in the task state, which the belief state would infer across time. Another example of state uncertainty is perceptual uncertainty. One study, published the same year as my paper covered in Chapter 1, used a belief state TD model to capture dopamine signals observed in primates during a perceptual decision-making task (Lak et al, 2017). In this task, a random dot motion stimulus was presented following a fixation cue (Nomoto et al, 2010). The coherence of the random dot motion stimulus varied from 50% (very easy) to 0% (impossible). Based on the perceived direction of the random dot motion (left or right), the animal would make an action that would result in a reward if the decision were correct. The authors modeled the belief state as a probability distribution over a range of motion directions and coherences, and used this to compute the value function and prediction errors at the onset of the random dot motion stimulus (Lak et al, 2017). It is important to note that in this perceptual decision-making task, the belief state is computed very differently than in the 90%-rewarded task described above. Rather than being yoked to elapsed time during a trial and probability of reward, the belief state is based on a noisy sampling of a perceptually challenging stimulus. Furthermore, the belief state is defined over an arbitrary large range (depending on how it is discretized in the model) of perceptual states, ranging from 50% leftward coherence to 50% rightward coherence, rather than just two states (ISI vs. ITI). This illustrates that the belief state TD model is a flexible framework that can incorporate diverse types of state uncertainty relevant to a particular task.
The computations necessary to calculate the belief state are diverse and highly specialized depending on the task. Given that the belief state TD model is a relatively young theory, the neural implementation of computing a belief state, and conveying this efficiently (regardless of the task) into the sub-cortical reinforcement learning circuitry in order to formulate value predictions, is not known. While the TD ‘feature representation’ has long been hypothesized to be computed by cortical structures (Houk et al, 1995), some recent observations cast doubt on this hypothesis. First, inactivation of the medial prefrontal cortex and the orbitofrontal cortex had only mild effects on dopamine signals during an operant conditioning task in rats (Jo and Mizumori, 2015; Jo at al, 2013). If a feature representation as basic as the CSC—needed to track time between cue and reward—were affected by cortical inactivation, one would have predicted a dramatically increased US response in these experiments. However, if the cortex is conveying a belief state representation (rather than a simple timing mechanism), it is possible that previously utilized experimental paradigms did not sufficiently implicate a belief state to observe an effect upon inactivation. Nonetheless, it is unclear from existing data whether cortex is truly needed for the dopamine system to compute prediction errors. Second, a recent study showed that monosynaptic inputs to dopaminergic neurons, from many different regions of the brain, themselves reflect pre-computed error signals (Tian et al, 2016). Therefore, the computation of prediction errors may be much more distributed than previously thought, making it appear unlikely that the neural correlates of reinforcement learning (including the feature representation) can be clearly mapped onto a particular brain structure. Despite these potential pitfalls, I show in Chapter 2 that the medial prefrontal cortex has a critical role in computing the belief state under conditions of state uncertainty. I directly compare the effect of inactivating the medial prefrontal cortex in a fully observable versus a partially observable task, and recapitulate my result using
the computational model developed in Chapter 1. While Chapter 1 tests the predictions of a reinforcement learning algorithm, Chapter 2 examines how this algorithm is structurally implemented in the brain. I argue that the prefrontal cortex has a specialized role in reinforcement learning, which is in line with Houk’s original suggestion that the cortex is responsible for conveying a feature representation. However, this role applies to a more specialized feature representation than previously thought.
CHAPTER 1

Dopamine Reward Prediction Errors Reflect Hidden State Inference Across Time

This chapter is based on a paper I published together with Professors Uchida and Gershman. I designed the behavioral tasks with Professor Uchida. I collected all data and performed all data analysis. I coded the computational models with help from Professor Gershman. I wrote the manuscript with Professors Uchida and Gershman.
INTRODUCTION

Midbrain dopamine neurons are thought to drive associative learning by signaling the difference between actual and expected reward, termed reward prediction error (RPE) (Schultz et al., 1997; Bayer and Glimcher, 2005; Cohen et al., 2012; Eshel et al., 2015). In particular, dopaminergic responses bear a striking resemblance to the error signal in a simple machine learning algorithm known as temporal difference (TD) learning (Schultz et al., 1997; Sutton and Barto, 1990). Several observations support this hypothesis (Schultz et al., 1997; Bayer and Glimcher, 2005; Cohen et al., 2012). 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, 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 (Gershman et al., 2010; Gershman et al., 2015). 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. While this idea has been explored theoretically (Daw et al, 2006; Rao, 2010), the empirical evidence remains sparse.

In the present study, 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, reward was delivered deterministically (100% rewarded). Our first task resembles other studies that examined dopamine signaling in tasks with variable ISIs (Fiorillo et al, 2008; Nomoto et al, 2010; Pasquereau and Turner, 2015). This previous work, particularly Fiorillo et al, 2008, described a mathematical framework for how temporal expectation influences dopamine RPEs. This work demonstrated that a hazard function, or a temporally blurred ‘subjective’ hazard function, describes temporal expectancy in the case of 100% reward delivery. Expanding upon this previous work, we also included a second task in which reward was occasionally omitted (90% rewarded). In the second task, the animal cannot initially be sure whether the absence of reward means that it was delayed or omitted entirely. As time elapses following cue onset, the animal’s belief that reward will arrive gradually yields to the belief that an omission trial occurred. Our results showed striking differences in dopamine signaling between 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 novel evidence that dopaminergic RPEs are shaped by state uncertainty.
RESULTS

Behavioral task and electrophysiology

We trained mice on either of two tasks (Figure 1.1a,b) (separate sets of 3 and 4 mice in Task 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 (mean: 2s, S.D.: 0.5s) defined over 9 timepoints ranging from 1.2s to 2.8s (Figure 1.1c,d). Odor B and C trials had constant ISIs of 1.2s and 2.8s, respectively. We included odors B and C to examine the effect of temporal delay on dopamine RPEs. On odor D trials, reward was never delivered. In a subset of mice trained on Task 2 (‘Task 2b’), we included an odor followed 2s later by

Figure 1.1 | Task design. (a) In Task 1, rewarded odors forecasted a 100% chance of reward delivery. Odors B and C trials had constant ISIs, while odor A trials had a variable ISI drawn from a discretized Gaussian distribution defined over 9 timepoints. (b) In Task 2, rewarded odors forecasted a 90% chance of reward delivery. ISIs for each odor were identical to Task 1. (c) Histogram of ISIs for odor A trials during an example Task 1 recording session, showing 9 possible reward delivery times. (d) Histogram of ISIs for odor A trials during an example Task 2 recording session. (e-f) Averaged non-normalized PSTH for licking behavior across all Task 1 (e) and Task 2 (f) recording sessions. Animals lick sooner for Odor B (ISI = 1.2s) than for Odor C (ISI = 2.8s) trials. Licking patterns for Odor A (variable ISI centered around 2.0s) fall in between licking patterns for Odor B and Odor C.
reward in order to compare dopamine RPEs in omission trials for constant versus variable ISIs. Mice learned to lick in anticipation of 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, 1-way ANOVA; $n = 9$ sessions, $F_{1,16} > 60$, $P < 1 \times 10^{-6}$ for odors A-C in Task 2, 1-way ANOVA; $n = 22$ sessions, $F_{1,42} > 100$, $P < 5 \times 10^{-13}$ for odors A-C in Task 2b, 1-way ANOVA). Anticipatory lick rate did not increase above baseline for odor D trials ($P > 0.10$ for all Tasks, odor D, 1-way ANOVA). (d) Average licking PSTH for all sessions across all animals trained on Task 1. (e) Average licking PSTH for all sessions across all animals trained on Task 2 (including Task 2b).

Figure 1.2 | Mice learn to lick in anticipation of reward. (a-c) Anticipatory lick rate increased above baseline for all rewarded trial types ($n = 26$ sessions, $F_{1,50} > 150$, $P < 6 \times 10^{-17}$ for odors A-C in Task 1, 1-way ANOVA; $n = 9$ sessions, $F_{1,16} > 60$, $P < 1 \times 10^{-6}$ for odors A-C in Task 2, 1-way ANOVA; $n = 22$ sessions, $F_{1,42} > 100$, $P < 5 \times 10^{-13}$ for odors A-C in Task 2b, 1-way ANOVA). Anticipatory lick rate did not increase above baseline for odor D trials ($P > 0.10$ for all Tasks, odor D, 1-way ANOVA). (d) Average licking PSTH for all sessions across all animals trained on Task 1. (e) Average licking PSTH for all sessions across all animals trained on Task 2 (including Task 2b). Mice ramped up their licking rates sooner and more steeply for odor B (ISI = 1.2s) over odor C (ISI = 2.8s) (Figure 1.1e,f), and for 100% rewarded odors over 90% rewarded odors (Figure 1.1e,f, Figure 1.2d,e). In both Tasks 1 and 2, licking patterns for odor A trials (average ISI = 2.0s) fell in between licking patterns for odor B and odor C. Lick rates following odor D, which never predicted reward, did not change significantly from baseline ($F < 2.5$, $P > 0.10$ for both tasks, 1-way ANOVA), demonstrating that mice learned the odor-outcome association.
We recorded the spiking activity of neurons in the VTA (387 neurons in 7 animals, see Figure 1.3 for recording sites) while animals performed Task 1 or 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 optic fiber positioned near our electrodes, and classified units as dopaminergic when they responded to light reliably with short latency (Figure 1.4; see Methods) (Cohen et al, 2012; Eshel et al, 2015; Tian et al, 2015).

**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 (Figure 1.5). 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 ($n = 30$).
Figure 1.5 | Averaged dopamine activity in Tasks 1 and 2 shows different patterns of modulation over variable ISI interval. (a) Average non-normalized PSTH for all 30 dopamine neurons recorded during Odor A trials in Task 1. Average pre- and post-reward dopamine RPE’s were negatively modulated by time (post-reward firing: $F_{8.232} = 5.56, P = 1.9 \times 10^{-6}$, 2-way ANOVA; factors: ISI, neuron; pre-reward firing: $F_{8.232} = 4.76, P = 2.0 \times 10^{-5}$, 2-way ANOVA; factors: ISI, neuron). (b) Average PSTH for all 43 dopamine neurons recorded during Odor A trials in Task 2 (includes neurons from Task 2b). Pre-reward dopamine RPE’s (400-0ms prior to reward onset) tended to be negatively modulated by time, while post-reward RPE’s (50-300ms following reward onset) tended to be positively modulated by time (post-reward firing: $F_{8.336} = 8.23, P = 3.48 \times 10^{-10}$, 2-way ANOVA; factors: ISI, neuron; pre-reward firing: $F_{8.336} = 7.86, P = 1.0 \times 10^{-9}$, 2-way ANOVA; factors: ISI, neuron). (c-f) Average PSTHs for odor B and C trials in Tasks 1 and 2. (g-h) Summary plots for average pre- and post-reward firing (mean ± s.e.m.).
Our quantification of post-reward firing revealed that, on average, post-reward firing was modulated negatively by time (Figure 1.5g). 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 ($F_{8,232}=4.76$, $P = 2.0 \times 10^{-5}$, 2-way ANOVA; factors: ISI, neuron). We computed pre-reward firing as the firing rate 400-0ms prior to reward onset, minus the firing rate 1000-0ms prior to odor onset. We found that pre-reward firing mirrored the post-reward pattern of negative modulation by time (Figure 1.5a,g). Therefore, in the case of 100% reward probability, both pre- and post-reward dopamine firing decreased as a function of time. This result is 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) (Fiorillo et al, 2008; Nomoto et al, 2010; Pasquereau and Turner, 2015).

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- and post-reward firing in Task 2 was calculated as described above for Task 1. We found that post-reward firing baseline (Hamid et al, 2016). Our quantification of post-reward firing revealed that, on average, post-reward firing was modulated negatively by time (Figure 1.5g). 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 ($F_{8,232}=4.76$, $P = 2.0 \times 10^{-5}$, 2-way ANOVA; factors: ISI, neuron). We computed pre-reward firing as the firing rate 400-0ms prior to reward onset, minus the firing rate 1000-0ms prior to odor onset. We found that pre-reward firing mirrored the post-reward pattern of negative modulation by time (Figure 1.5a,g). Therefore, in the case of 100% reward probability, both pre- and post-reward dopamine firing decreased as a function of time. This result is 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) (Fiorillo et al, 2008; Nomoto et al, 2010; Pasquereau and Turner, 2015).

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- and post-reward firing in Task 2 was calculated as described above for Task 1. We found that post-reward firing baseline (Hamid et al, 2016). Our quantification of post-reward firing revealed that, on average, post-reward firing was modulated negatively by time (Figure 1.5g). 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 ($F_{8,232}=4.76$, $P = 2.0 \times 10^{-5}$, 2-way ANOVA; factors: ISI, neuron). We computed pre-reward firing as the firing rate 400-0ms prior to reward onset, minus the firing rate 1000-0ms prior to odor onset. We found that pre-reward firing mirrored the post-reward pattern of negative modulation by time (Figure 1.5a,g). Therefore, in the case of 100% reward probability, both pre- and post-reward dopamine firing decreased as a function of time. This result is 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) (Fiorillo et al, 2008; Nomoto et al, 2010; Pasquereau and Turner, 2015).
was significantly modulated by ISI length \((n = 43\) neurons, \(F_{8,336} = 8.23, P = 3.48 \times 10^{-10},\) 2-way ANOVA; factors: ISI, neuron). Strikingly, we observed the opposite trend of modulation over time, compared to Task 1. On average, reward delivery elicited a phasic response that was smallest for shorter ISIs and greatest for the longest ISI (Figure 1.5b,h). Pre-reward firing in Task 2 was also significantly modulated by ISI length \((F_{8,336} = 7.86, P = 1.0 \times 10^{-9},\) 2-way ANOVA; factors: ISI, neuron), and tended to decrease throughout the variable ISI interval (Figure 1.5b,h). In sum, 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 ISI on every trial and drew a best-fit line through the data (Figure 1.7a-f). Based on the slopes of these best-fit lines, we found that 23/30 neurons tended towards negative modulation by time in Task 1 (95% CI < 0 for 11/30 neurons; Figure 1.7 | Individual dopamine neurons show opposing patterns of post-reward firing in Tasks 1 and 2. (a,b) PSTH for two example dopamine neurons during odor A trials of 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 (50-300ms following reward onset) for each odor A trial in Task 1 (e) or Task 2 (f). (g,h) Slopes of best-fit lines in Task 1 (g) or Task 2 (h), as shown in (e) and (f), for all dopamine neurons recorded. Purple shading indicates \(P < 0.05,\) or a 95% confidence interval for the slope coefficient that does overlap with 0.
Figure 1.7g). 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; Figure 1.7h). We repeated the same analysis for pre-reward firing in both Tasks. In Task 1, pre-reward firing for 19/30 individual neurons tended towards negative modulation by time (95% CI < 0 for 14/30 neurons). In Task 2, pre-reward firing for 32/43 neurons tended towards negative modulation by time (95% CI < 0 for 9/43 neurons). Therefore, individual neurons recorded in each Task tended to reflect the trends of temporal modulation described above.

To summarize, in Tasks 1 and 2, we found that dopaminergic RPEs were modulated over time for various ISI lengths. Pre-reward firing in both tasks tended to decline throughout the variable ISI interval. However, post-reward firing showed opposite trends of temporal modulation in these two tasks. In Task 1, post-reward firing showed negative temporal modulation, and 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 (Fiorillo et al, 2008; Kobayashi and Schultz, 2008; Jo and Mizumori, 2015). 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, we would expect the difference between post-reward firing for odor B trials (ISI = 1.2s) and odor C trials (ISI = 2.8s) to account for the difference between post-reward firing for the earliest and latest rewards for odor A trials (ISI = 1.2s and 2.8s, respectively; Figure 1.5h). In Task 2, we found that the
average post-reward firing rate for odor C was about 1Hz higher than for odor B. This modest difference was not significant (n = 14 neurons; $F_{1,13} = 0.85$, $P = 0.37$, 2-way ANOVA; factors: odor, neuron). Moreover, the latest possible reward delivery following odor A (ISI = 2.8s) elicited post-reward firing significantly higher than odor C post-reward firing (n = 14 neurons; $F_{1,13} = 7.15$, $P = 2 \times 10^{-2}$, 2-way ANOVA; factors: odor, neuron). These results indicate that the positive temporal modulation of post-reward firing observed in Task 2 cannot 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 (Schultz et al, 1997). We therefore examined whether previously proposed TD learning models can 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 (Sutton, 1988):

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

where $E[\cdot]$ denotes an average over randomness in reward delivery, and $\gamma$ is a discount factor that down-weights future rewards. The goal of reinforcement learning models is to learn correct value estimates so as to maximize future rewards.
The original application of TD learning to the dopamine system (Schultz et al, 1997) assumed a “complete serial compound” (CSC) representation $x(t) = \{x_1(t), x_2(t), \cdots \}$ as stimulus features for value computation (Figure 1.8a). The onset of a reward-predictive stimulus initiates a ballistic sequence of sub-states marking 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:

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

(2)

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

$$\Delta w_i = \alpha x_i(t) \delta(t)$$

(3)

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

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

(4)

We first tested whether TD learning with the CSC can 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 (Figure 1.8b). 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 cannot explain our data.
We next asked whether a simple modification of the original model could better account for our results: a ‘reset’ feature that sets the RPE to zero after reward arrives (Suri and Schultz, 1998; Suri and Schultz, 1999). This model rectifies one key inconsistency between 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 (Hollerman and Schultz, 1998). 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, that is, reward expectation that grows over time, increasingly suppressing excitation towards the end of the variable ISI interval (Figure 1.8c). A pattern of decreasing RPEs over time matches our Task 1 data. However, in Task 2, this model also produced RPEs that generally decreased throughout the variable ISI interval, deviating from the trend of our data. Therefore, this proposed modification to the original model cannot explain our results. Our data do not completely rule out other reset devices, such as resetting the stimulus trace following reward (Brown et al, 1999). However, as pointed out by Daw and colleagues (Daw et al, 2006), such a reset device assumes

---

**Figure 1.8** | TD with CSC model, with or without Reset, is inconsistent with our data. (a) Schematic adapted from Schultz et al, 1997. 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. (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.
an inferred state change, and may not generalize gracefully 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 animal in the above tasks is knowing whether it is in one of the two states: the ISI state during which the animal expects reward, and the ITI state during which no reward is expected. In Task 1, these two states are fully observable, since cue onset unambiguously signals a transition to the ISI state, and reward onset unambiguously signals a transition to the ITI state; no transitions occur without one of these events (Figure 1.9a). Thus, the states are fully observable, and the only computational

![Figure 1.9](image)

**Figure 1.9 | Belief state model parameters (see Computational Methods for details).** (a,b) Schematic of the difference between modeling Task 1 and 2. Whereas cue can never lead back to the ITI sub-state in Task 1, this ‘hidden’ transition occurs in 10% of trials (omission trials) in Task 2. (c,d) Observation matrices for Task 1 and 2. In a tiny percentage of cases in Task 2, the animal observes ‘cue’ when in fact it transitions from the ITI back to the ITI during an omission trial. This occurs in such a small proportion of sub-state 15→15 transitions that it cannot be visualized in this map, so we have indicated values that differ between Tasks directly below the matrices. See Methods for how these values were computed. (e,f) Transition matrices for Task 1 and 2. The probability of transitioning from the sub-state 15 back to the sub-state 15 is slightly higher in Task 2, due to the presence of omission trials. The baseline probability of transitioning out of the ITI in either task is so small that it cannot be visualized here, so we have indicated values that differ between Tasks directly below the matrices.
problem is predicting reward. In Task 2, omission trials cause the ITI state to self-transition (while still emitting a cue). This means that both ITI-to-ISI and ITI-to-ITI generate the same observation, rendering the states partially observable or “hidden” (Figure 1.9b). 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 ITI or ISI at a given moment. To incorporate this process in our model, here we assume that the ISI and ITI comprise temporal “sub-states” with the analogy to the CSC model (Figure 1.10a,b). The normative solution to the hidden state inference problem is given by Bayes’ rule, which stipulates how an animal’s probabilistic beliefs about states should be updated over time:

\[
b_i(t + 1) \propto p(o(t)|i) \sum_j p(i|j) b_j(t - 1)
\]

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 \( o(t) \in \{\text{cue}, \text{reward}, \text{null}\} \) under hypothetical sub-state \( i \), and \( p(i|j) \) is the probability of transitioning from sub-state \( j \) to sub-state \( i \) (Figure 1.9c-f).
The vector $b(t)$ functions as a “belief state” that can substitute for the CSC in learning equations (2) and (3) shown above (Figure 1.10c,d). In Task 1, where the observations are unambiguous, the belief state is identical to the CSC. In Task 2, the belief state departs 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).

**Figure 1.10 | Belief state model is consistent with our data.** (a,b) In our model, the ISI and ITI states comprise sub-states 1-15 (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) Belief state model captures the opposing post-reward firing patterns between Task 1 (e) and Task 2 (f). This model also captures negative temporal modulation of pre-reward firing in both Tasks.
While 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 (Daw et al., 2006), where sub-states are replaced by dwell-time distributions in each state. These models are mathematically equivalent; we chose the sub-state formulation in order 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 (Figure 1.10e,f). In Task 1, the states are fully

Figure 1.11 | Belief state model shapes value signals that differ between Tasks 1 and 2, leading to opposite patterns of post-reward modulation over time. (a,b) As time elapses following odor onset in Task 1, the belief state proceeds through ISI sub-states (i1-i14) 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) As time elapses following odor onset in Task 2, the belief state comprises a probability distribution that gradually decreases for ISI sub-states (i1-i14) and gradually increases for the ITI sub-state (i15). This produces a value signal that declines for longer ISIs, resulting in the least suppression of $\delta(t)$ for the latest ISI.
observable and thus the belief state is uniform throughout the variable ISI interval (Figure 1.10c, Figure 1.11a,b): as soon as the cue comes on, the belief state encodes 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 reward is greater at later ISIs than at sooner ISIs, later sub-states accrue higher weights than earlier sub-states, producing a ramping value signal (Figure 1.11b). This ramping value signal results in RPEs that are increasingly suppressed towards the end of the variable ISI interval (Figure 1.10c, Figure 1.11a,b; see Figure 1.12a,c for quantification), producing a pattern of negative modulation by time similar to 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 interval. As time elapses and reward fails to arrive, the belief state progressively shifts in favor of the ITI over the ISI (Figure 1.10d, Figure 1.11d). Rewards sometimes arrive at the latest ISIs, increasing the weights for the corresponding sub-states. However, the belief state for these late timepoints is so skewed towards the ITI that the value signal actually decreases relative
to earlier timepoints (Figure 1.11c,d). This decreasing value signal results in pre-reward RPEs that are most suppressed, and post-reward RPEs that are least suppressed, at the end of the variable ISI interval (Figure 1.10f, Figure 1.11c,d; see Figure 1.12b,d for quantification). Post-reward RPEs towards the end of the interval were nearly as large as unpredicted rewards, both in our model results (Figure 1.12d) and our data (Figure 1.13b). Therefore, our model captures the pattern of pre- and post-reward RPEs in our Task 2 data. Importantly, the belief state model captures the striking 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 Odor A omission trials (2s variable ISI), we found that the trough of the dip in dopamine firing occurred slightly later than the trough for Odor B trials (2s constant ISI). This shift in the trough of the omission response was also reproduced by our belief state TD model (Figure 1.14).

Figure 1.13 | Post-reward RPEs normalized to free water response. In order to compare the magnitude of dopamine RPEs across different populations of neurons, post-reward RPEs must be normalized to the free water response (see Eshel et al, 2016). (a) Post-reward RPEs in Task 1 are suppressed to about 60% of the free water response at the beginning of the interval (1.2s), and are suppressed to about 40% of the free water response at the end of the interval (2.8s). RPEs plotted are from the 29 neurons for which we included 2-5 free water deliveries during recording. (b) Post-reward RPEs in Task 2 are suppressed to about 60% of the free water response at the beginning of the interval (1.2s), and are suppressed to about 90% of the free water response at the end of the interval (2.8s). RPEs plotted are from the 34 neurons for which we included 2-5 free water deliveries during recording. (c) Schematic of data in (a) and (b), highlighting that the opposing trends of temporal modulation in Tasks 1 and 2 arise mostly from diverging patterns towards the end of the interval.
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 of 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 (Figure 1.15). Therefore, our modeling result is relatively agnostic to the precise shape of the learned ISI distribution.

Finally, while our model captures the trends of pre- and post-reward temporal modulation in both Tasks 1 and 2, the overall dopamine firing in Task 1 is much larger than predicted by our model. What could cause the discrepancy in post-reward RPE magnitude between our Task 1 data and model? Because our mice are trained on an odor-outcome association, the exact time when the animals sniff and detect ‘odor ON’ is jittered from trial to trial. This temporal jitter limits how precisely the animal can anticipate reward timing. Therefore, because our model does not incorporate this trial-by-trial jitter, it suppresses RPEs more effectively, particularly in Task 1 conditions that allow reward timing to be predicted perfectly by the end of the interval.
Furthermore, our mice are trained for a relatively short length of time (~1-2 weeks) prior to recording, potentially limiting the extent to which RPEs can be suppressed. Indeed, training our model on fewer trials increases the magnitude of post-reward RPE’s.

**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 (Oswal et al, 2007; Janssen and Shadlen, 2005; Tsunoda and Kakei, 2008; Ghose and Maunsell, 2002). 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 other studies that analyzed dopaminergic RPEs in tasks with variable ISIs (Fiorillo et al, 2008; Nomoto et al, 2008; Pasquereau and Turner, 2015), the variably timed event always occurred (100% event probability) and the ISI was drawn from a uniform distribution. With respect to both pre- and post-reward dopamine firing, all of these studies found a pattern of decreasing excitation over elapsed time, thought to correspond to a rising hazard function that increasingly suppressed later RPEs. However, one aspect of previous work could not be explained 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 (Fiorillo et al., 2008).

Intriguingly, when we trained our belief state TD model on an exponential distribution similar to this previous work, our model was able to reproduce the negative temporal modulation of post-reward RPEs (Figure 1.16d,e).

Our data in Task 1, which utilized a Gaussian ISI distribution and 100% reward probability, also revealed a pattern of decreasing pre- and post-reward dopamine firing, which matches the proposal that a hazard function may describe the trend of temporal expectancy reflected by dopamine RPEs (Figure 1.17). However, our data in Task 2 cannot be explained by a hazard function, nor can they be explained by a temporally blurred subjective hazard function, computed by blurring the probability distribution function with a Gaussian whose standard deviation scales with elapsed time (see Janssen and Shadlen, 2005; Tsunoda and Kakei, 2008; Methods) (Figure 1.17). Plotting the hazard function and subjective hazard functions for Task 2 reveals that both of these functions find a minimum for the earliest rewards. However, our data
indicates that temporal expectation is at its maximum for the earliest rewards, because the earliest post-reward RPE’s are most suppressed (Figure 1.5b). We illustrated this contrast by plotting the value function from our belief state TD model alongside the hazard function for Task 2 (Figure 1.17a). In sum, a hazard function may describe temporal expectancy for 100% rewarded conditions. However, temporal expectancy is dramatically altered in conditions involving uncertainty about whether the event will occur at all.

**Figure 1.17** | **Hazard and subjective hazard functions cannot explain the trend of our data.** (a) Hazard and subjective hazard 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 1.11 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.
DISCUSSION

In this work, we examined how dopaminergic RPE signals change with respect to reward timing and probability. Our experimental results showed that, depending on whether or not reward is 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 (Daw et al, 2006). 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 were classically represented as a cascade of sub-states (the “complete serial compound” or CSC) that track elapsed time following stimulus onset (Schultz et al, 1997; Sutton and Barto, 1990). 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 passively observing its environment (Friston, 2005; Lee and Mumford, 2003). 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 pawprint. By conditioning its belief state on observations of its environment, the antelope may predict the lion’s presence. Following earlier theoretical work (Daw et al, 2006; Rao, 2010; Kakade and Dayan, 2002) we argue that this inferential process is at play 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.
This work follows two recent empirical studies that explored a state-based framework in the striatum and the VTA (Stalnaker et al, 2016; Takahashi et al, 2016). 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 (Stalnaker et al, 2016). In the second of these studies, a state-based model was used to capture the effect of a striatal lesion, which selectively impacted the temporal specificity of dopaminergic prediction errors but spared value-related prediction errors (Takahashi et al, 2016). These two studies support our claim that a belief state representation may be at play in the basal ganglia reward-processing circuitry.

Previous studies have shown a pattern of decreasing RPE’s over time during tasks in which ISIs are drawn from a uniform probability distribution (Fiorillo et al, 2008; Nomoto et al, 2010; Pasquereau and Turner, 2015). Can our model account for the temporal modulation of dopamine RPEs in these previous studies? Upon training the belief state TD model on a uniform distribution of reward timings, our model elicited negative temporal modulation of RPE signals (Figure 1.15), 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 RPE’s over time. TD learning using the complete serial compound and reset also produced a pattern of decreasing excitation over time when trained on a flat probability distribution (Figure 1.16a-c). 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 in both of these task conditions allowed us to distinguish between
the predictions of various associative learning models, thereby expanding upon 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 (Fiorillo et al, 2008; Kobayashi and Schultz, 2008; Jo and Mizumori, 2015). These works showed that dopamine RPEs are less suppressed for lengthier ISIs, likely due to scalar timing uncertainty. For simplicity, our belief state model omitted the effect of temporal uncertainty in order 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 standard deviation is proportional to elapsed time (Takahashi et al, 2016; see Figure 1.18). To create this ‘blurred’ belief state model, we fit a scalar timing noise parameter to account for post-reward RPEs for 1.2s and 2.8s constant delays (Odors B and C). This temporally blurred belief state model still

Figure 1.18 | Belief state model can be adapted to incorporate temporal uncertainty. (a) Belief state at time t was blurred by a Gaussian distribution with standard deviation proportional to t, in order to capture temporal uncertainty. (b,c) Belief state model with temporal uncertainty captures our data well. Post-reward RPEs are less suppressed for Odor C than for Odor B in both tasks, due to temporal uncertainty that grows with elapsed time.
captured our data well in Tasks 1 and 2.

Although we have focused on the belief state TD model, another prominent account replaces the CSC with “microstimulus” features—temporally diffuse versions of the discrete time markers in the CSC (Ludvig et al, 2008). 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 standard deviation increases and amplitude decreases with the post-stimulus interval. Although the microstimulus and belief state models are typically thought of as alternatives [see Gershman et al, 2014, for a review], they can be conceived as realizations of the same idea at different levels of analysis.

Examining the belief state over time, we can see that the posterior over each sub-state peaks at a specific moment during the trial (Figure 1.11). In Task 2, the peaks become progressively lower as a function of time, due to the increased probability of a state transition. This decrease in amplitude mirrors 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 (Mello et al, 2015). 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. While belief states are cognitive abstractions, they could be realized in the brain by neurons with temporal receptive field structure resembling microstimuli.
METHODS

Animals. We used 7 adult male mice, heterozygous for Cre recombinase under the control of the DAT promoter (B6.SJL-Slc6a3tm1.1(cre)Bkmm/J, The Jackson Laboratory) and backcrossed for >5 generations with C57/BL6J mice (Backman et al, 2007). 3 animals were used in Task 1 (Figure 1.1a), 1 animal was used in Task 2 (Figure 1.1b), and 3 animals were used in Task 2b (Figure 1.14a). Animals were housed on a 12-h dark/12-h light cycle (dark from 7AM to 7PM). We trained animals 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 and approved by the Harvard Institutional Animal Care and Use Committee.

Surgery and viral injections. We performed all surgeries under aseptic conditions with animals under isoflurane (1-2% at 0.5-1.0L/min) anesthesia. Analgesia (buprenorphine, 0.1mg/kg, intraperitoneal) was administered pre-operatively and at 12-h checkpoints post-operatively. We performed two surgeries, both stereotactically targeting the left VTA (from bregma: 3.1mm posterior, 0.6mm lateral, 4.2mm ventral). In the first surgery, we injected 500nL of adeno-associated virus (AAV, serotype 5) carrying an inverted ChR2 (H134R) fused to the fluorescent reporter eYFP and flanked by double loxP sites (Cohen et al, 2012; Atasoy et al, 2008). We previously showed that the expression of this virus is highly selective and efficient in dopamine neurons (Cohen et al, 2012). 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 pre-restriction body weight. We habituated and briefly head-retrained mice for 2-3 days before training. Odors were delivered to animals with a custom-made olfactometer (Uchida and Mainen, 2003). Each odor was dissolved in mineral oil at 1/10 dilution. 30uL of diluted odor was placed into glass fiber filter-paper, and then diluted with filtered air 1:20 to produce a total 1L/min flow rate. Odors 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 animals. We automatically detected licks by measuring breaks of an infrared beam 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 9 points ([1.2s 1.4s 1.6s 1.8s 2.0s 2.2s 2.4s 2.6s 2.8s]; mean = 2s; SD = 0.5s), prior to reward delivery. For both Tasks 1 and 2, rewarded odor B and odor C trials consisted of 1s odor presentation followed by either 1.2s or 2.8s delay from odor onset, respectively, prior to reward delivery (Figure 1.1a,b). In Task 2b, rewarded odor B trials consisted of 1s odor presentation followed by 2s delay from odor onset; odor C was not given (Figure 1.14a). 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 tasks, reward size was kept constant at 3uL. Trial type was drawn pseudorandomly from a scrambled array of trial types, in order to keep the proportion of trial types constant between sessions. The ITI between trials was drawn from an exponential distribution (mean = 12-14s) in order to ensure a flat hazard function. Animals performed between 150-300 trials per session.
Electrophysiology. We based recording techniques on previous studies (Cohen et al, 2012; Eshel et al, 2015; Tian and Uchida, 2015). We recorded extracellularly from the VTA using a custom-built, screw-driven Microdrive (Sandvik, Palm Coast, Florida) containing 8 tetrodes glued to a 200µm optic fiber (ThorLabs). Tetrodes were glued to the fiber and clipped so that their tips extended 200-500µm 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 9000 Hz and recorded continuously at 32kHz. To extract spike timing, signals were band-pass-filtered between 300 and 6000Hz and sorted offline using MClust-3.5 (A.D. Redish). At the end of each session, the fiber and tetrodes were lowered by 75um to record new units the next day. To be included in the dataset, a neuron had to be well-isolated (L-ratio < 0.05; Schmitzer-Torbert et al, 2005) and recorded within 300um of a light-identified dopamine neuron (see below) to ensure that it was recorded in the VTA. We also histologically verified recording sites by creating electrolytic lesions using 10-15s of 30µA direct current.

To unambiguously identify dopamine neurons, we used ChR2 to observe laser-triggered spikes (Cohen et al, 2012; Lima et al, 2009; Kvitsiani et al, 2013). 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 10 473nm light pulses, each 5ms long, at 1, 5, 10, 20, and 50Hz, with an intensity of 5-20mW/mm² at the tip of the fiber. Spike shape was measured using a broadband signal (0.1-9,000Hz)
sampled at 32kHz. To be included in our dataset, neurons had to fulfill 3 criteria (Cohen et al, 2012; Eshel et al, 2015; Tian and Uchida, 2015):

1) Neurons’ spike timing must be significantly modulated by light pulses. We tested this by using the Stimulus-Associated spike Latency Test (SALT) (Kvitsiani et al, 2013). We used a significance value of $P < 0.05$, and a time window of 10ms after laser onset.

2) Laser-evoked spikes must be near-identical to spontaneous spikes. This ensured that light-evoked spikes reflect actual spikes instead of photochemical artifacts. All light-identified dopamine neurons had correlation coefficients $> 0.9$ (Figure 1.4b,g).

3) Neurons must have a short latency to spike following laser pulses, and little jitter in spike latency (Figure 1.4 c,e,f). While others have used a latency criteria of 5ms or less (‘short latency’) (Cohen et al, 2012; Eshel et al, 2015; Tian and Uchida, 2015), we found that the high laser intensity required to elicit this short latency spike sometimes created a mismatched waveform, due to 2 neurons near the same tetrode being simultaneously activated. For this reason, we often decreased the laser intensity and elicited a spike 5-10ms (‘longer latency’) after laser onset. We separately analyzed neurons in both the ‘short latency’ and ‘longer latency’ categories, and found qualitatively similar results in each group. Therefore, we pooled all dopamine neurons with latencies below 10ms 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 1ms bins. Averaged
PSTHs shown in figures were smoothed with a box filter of 100-150ms. Average pre-reward firing rates were calculated by counting the number of spikes 0-400ms prior to reward onset. We also attempted using window sizes ranging from 200-500ms, and these produced similar results. Average post-reward firing rates were calculated by counting the number of spikes 50-300ms after reward onset in both Tasks 1 and 2. Both pre- and post-reward responses were baseline-subtracted, with baseline taken 0-1s prior to odor onset.

We plotted a subjective hazard rate by blurring the probability distribution function $p(t)$ by a normal distribution whose standard deviation scales with elapsed time. Similar to previous work (Janssen and Shadlen, 2005), we used a Weber fraction $\phi = 0.25$:

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

**Statistics.** No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications (Cohen et al, 2012; Eshel et al, 2015; Tian and Uchida, 2015; Tsunoda and Kakei, 2008). Data collection and analysis were not performed blind to the conditions of the experiments. Animals were chosen at random for Tasks 1 or 2. All trial types were randomly interleaved within a single recording session. We verified that all groups of data (including both electrophysiology and behavior) 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 2-factor ANOVA, with neuron and ISI as factor. To test whether licking was modulated by odor identity, we used a 1-factor ANOVA, with Odor identity as a factor. To test whether individual neurons’ RPEs were modulated by factors such as ISI length or lick rates, we fit a line to the data (dopamine RPEs...
versus ISI) and reported the slope. We also displayed a summary of whether or not the 95% confidence interval of the slope included 0 (shaded in Figure 1g,h).

**Code Availability.** Code used to implement the computational modeling in this manuscript can be found in a Supplementary Software section and at this GitHub link: 

https://github.com/cstarkweather

**Immunohistochemistry.** After 4-8 weeks of recording, we injected mice with an overdose of ketamine/medetomidine. Mice were exsanguinated with saline and perfused with 4% paraformaldehyde. We cut brains in 100um coronal sections on a vibrotome and immunostained with antibodies to tyrosine hydroxylase (AB152, 1:1000, Millipore) in order 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 AAV.

**Computational modeling.**

**Temporal difference (TD) Model:** We first simulated TD error signaling in our Tasks by using Temporal Difference learning with a complete serial compound representation, identical to the algorithm presented by Schultz and colleagues (Schultz et al, 1997). We set stimulus onset at $t = 20$, and set 9 possible reward times at $t = 26, 27, 28, 29, 30, 31, 32, 33, 34$. 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 10x simulations of each task,
with 5000 trials per simulation. The ‘TD with reset’ variant was simulated by setting the error term to 0 at any timesteps 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 Rao (Daw et al, 2006; Rao, 2010). To capture the discrete dwell times in our Tasks (1s 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 2.8s after odor onset), we coded a Markov equivalent of a Semi-Markov model (see Daw et al, 2006). This Markov equivalent contained 15 total hidden sub-states (see Figure 1.9a,b). Sub-states 1-5 corresponded to the passage of time during the 1s odor presentation; sub-states 6-14 corresponded to the passage of time preceding the 9 possible reward delivery time. Sub-state 15 corresponded to the ITI. If reward was received at the earliest possible time (1.2s), 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 (2.8s), 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 9 discrete timepoints. For example, the hazard rate of receiving reward at 1.2s would correspond to T(6,15), or the probability of transitioning from sub-state 6→15. 1 minus the hazard rate of receiving reward at 1.2s 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) to 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. 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:

$$T(15,15) = 1 - ITI_{hazard}$$
$$T(15,1) = ITI_{hazard}$$

Task 2:

$$T(15,15) = 1 - ITI_{hazard} \times 0.9$$
$$T(15,1) = ITI_{hazard} \times 0.9$$

This difference in $T$ between Task 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 3 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:

$$O(i,j,1) = \text{likelihood of observation of ‘null’, given } i \rightarrow j \text{ transition}$$
$$O(i,j,2) = \text{likelihood of observation of ‘cue’, given } i \rightarrow j \text{ transition}$$
$$O(i,j,3) = \text{likelihood of observation of ‘reward’, given } i \rightarrow j \text{ transition}$$

In order to switch from sub-state 15 (ITI) to sub-state 1 (first state of ISI), the animal must have an observation of the cue: $O(15,1,2) = 1$. In order 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:
\[ O(15,15,1) = 1 - ITI\_hazard\times0.1 \] (null observation)
\[ O(15,15,2) = ITI\_hazard\times0.1 \] (cue in a small percentage of cases)

This difference in O between Task 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 5000 trials each. We found that the model yielded asymptotic results after about 1000 trials. For this reason, the results shown in the text are taken from trials 2000-5000. In all simulations, we used a learning rate of \( \alpha = 0.1 \) and a discount factor of \( \gamma = 0.98 \).
CHAPTER 2

Medial Prefrontal Cortex Shapes Dopamine Reward Prediction Errors Under State Uncertainty

This chapter is based on a paper I submitted together with Professors Uchida and Gershman. I designed the behavioral tasks with Professor Uchida. I collected all data and performed all data analysis. I coded the computational models. I wrote the manuscript with Professors Uchida and Gershman.
INTRODUCTION

The ability to predict future outcomes is at the core of adaptive behaviors. In reinforcement learning theories, future outcomes are predicted based on the ‘state’ of the world defined by a set of information including the location of the animal, what objects are present, and elapsed time from certain events. A challenge in making predictions in natural environments is that the cues required to define the current state are often ambiguous, and it is difficult to know what state the animal is in in the first place. That is, the current state is ‘hidden’, and needs to be inferred from partial information (Courville et al, 2006; Gershman et al, 2010; Gershman et al, 2015). It has been proposed that, in the presence of such uncertainty, the brain computes future expectations based on a probability distribution defined over possible hidden states (a ‘belief state’) (Daw et al, 2006; Rao, 2010; Starkweather et al, 2017). Although theoretical and empirical evidence has begun to support this idea as it applies to the midbrain dopamine system (Rao, 2010; Lak et al, 2017), the neural mechanisms underlying hidden state inference remain largely unknown.

The activity of dopamine neurons is sensitive to reward expectation. It has been shown that dopamine neurons report a reward prediction error (RPE) signal that is thought to reflect the discrepancy between actual and predicted value (Schultz et al, 1997; Bayer et al, 2005; Cohen et al, 2012; Eshel et al, 2015). Dopamine neurons’ responses to reward-predictive cues scale with expected future reward (Fiorillo, Tobler, and Schultz, 2003; Cohen et al, 2012; Tian and Uchida, 2015). More importantly, dopamine reward responses are suppressed according to the magnitude of reward expectation (Fiorillo, Tobler, and Schultz, 2003; Cohen et al, 2012; Tian and Uchida, 2015). Previous studies have shown that the magnitude of dopamine reward responses is indeed
modulated by the moment-by-moment strength of reward expectation when the timing of reward is varied (Fiorillo, Newsome, and Schultz, 2008; Nomoto et al, 2010; Pasquereau and Turner, 2015). For instance, dopamine reward responses decrease as time elapses, as if reward expectation increases as a function of elapsed time. This result is consistent with the idea that reward expectation grows with the hazard rate (i.e. the likelihood of an event happening at a moment, given that the event has not happened yet). Notably, these prior studies with variable reward delivery times have utilized

Figure 2.1 | Classical conditioning tasks that vary reward timing produce divergent patterns of dopamine responses, depending on whether reward is delivered deterministically. (a) In Task 1, rewarded odors forecasted a 100% chance of reward delivery. The time between cue and reward (interstimulus interval, or ‘ISI’) was varied across trials. (b) The task ‘state’—here meaning ISI (reward will be delivered) or ITI (reward will not be delivered)—is fully observable in Task 1, because it is reliably signaled by sensory cues such as cue and reward. (c) The Task 1 belief state is fixed with 100% probability in the ISI state after cue onset. (d) RPEs decrease as a function of time when reward timing is varied under a 100% rewarded contingency. (e) In Task 2, rewarded odors forecasted a 90% chance of reward delivery. Similar to Task 1, the time between cue and reward (interstimulus interval, or ‘ISI’) was varied across trials. (f) The task state is hidden in Task 2, because it is not reliably signaled by cue onset (cue could lead back to the ITI, during which no reward will be delivered). (g) The Task 2 belief state is initially fixed with 90% probability in the ISI state after cue onset, but this probability gradually decreases as time elapses. Eventually, the belief state yields to the possibility that the trial will be unrewarded, allotting more probability to the unrewarded ITI state. (h) RPEs increase as a function of time when reward timing is varied under a 90% rewarded contingency.
experimental paradigms in which reward is always delivered. In contrast, a previous study showed that the hazard account does not hold in conditions in which reward is delivered in a probabilistic manner, and instead a model incorporating hidden state inference explains the data better under these conditions (Starkweather et al., 2017).

This previous study (Starkweather et al., 2017) recorded dopamine RPEs during a classical conditioning task in which reward timing was varied across trials (Figure 2.1). The activity of dopamine neurons exhibited distinct patterns of responses depending on whether reward was delivered in 100% of trials (Task 1) or 90% of trials (Task 2). In Task 1, dopamine reward responses were modulated negatively over time, as if expectation increased over time, consistent with the hazard rate account (Figure 2.1d). In a stark contrast, in Task 2, dopamine reward responses increased as time elapsed (Figure 2.1h), and the result could not be explained by a hazard rate. Computational modeling indicated that these divergent patterns of response are explained well by a reinforcement learning model in which reward expectation is computed over belief states. This model assumes transitions between two discrete states: the inter-stimulus interval (ISI) state during which reward is expected, and the inter-trial interval (ITI) state during which no reward is expected (Figure 2.1b,f). The animal infers which state it is in based on the presentation of odor cues, reward, and the elapsed time from these events. Importantly, probabilistic reward delivery renders the task states hidden: the animal cannot know for certain whether it is in the ISI state or ITI state (Figure 2.1f). Thus, in Task 2, after detecting the cue, the animal’s belief that it is in the ISI gradually yields to the belief that it is in the ITI (Figure 2.1g), resulting in RPEs that increased over elapsed time. In Task 1, the belief of being in the ISI is 100% after cue presentation (Figure 2.1c), thus, reward expectation grows with elapsed time.
and follows the hazard rate. Reward expectation computed over belief states was able to explain
the divergent response patterns in both tasks. These results demonstrated that in order to account
for reward expectations in these tasks, (1) even a simple classical conditioning paradigm must be
modeled with transitions between the ISI and ITI states (i.e. by explicitly modeling the ITI state),
and that (2) reward expectation is computed over the animal’s belief (inferred probability) over
these two possible states.

In this study, we sought to explore neural mechanisms underlying hidden state inference
using these two tasks. Specifically we sought to dissect the contribution of medial prefrontal
cortex (mPFC) to the reinforcement learning circuitry. Classically, reinforcement learning
models have postulated a cortical substrate for tracking the agent’s internal sense of time
(Schultz et al, 1997). In line with this theoretical prediction, previous studies in rodents
suggested that the mPFC is important in interval timing (Kim et al, 2009; Kim et al, 2013; Xu et
al, 2014). However, cortical inactivation studies have produced relatively mild effects on
dopamine RPEs (Jo, Lee, and Mizumori, 2013; Jo and Mizumori, 2015). Moreover, these studies
did not separate the contribution of interval timing judgment from hidden state inference.
Therefore, the involvement of the mPFC in reinforcement learning, as well as its role in hidden
state inference or interval timing, remains elusive. Our behavioral paradigms differentially
implicate both of these processes: hidden state inference comes into play only in Task 2, while
interval timing is needed to compute both the time-dependent ‘hazard’-like expectancy in Task 1
and the belief state in Task 2. We sought to test whether the neural substrate for hidden state
inference and interval timing can be separated, and whether mPFC regulates dopamine RPEs
through either (or both) of these two processes.
RESULTS

We trained animals on two classical conditioning tasks that produced differing patterns of dopamine RPEs across time, depending on whether or not reward was delivered with certainty. (Figure 2.1, Starkweather et al, 2017). In these tasks, odor cues predicted a delivery of reward with variable timing after an odor cue. The two tasks differed only with respect to whether reward was delivered in 100% of trials (Task 1) or 90% of trials (Task 2). Using these paradigms, we examined whether dopamine responses were affected by temporal inactivation of mPFC using pharmacogenetic inactivation.

KORD inactivates mPFC neurons

We first examined the efficacy and the time course of mPFC inactivation. We injected an adeno-associated virus (AAV) carrying kappa opioid receptor-based designer receptors exclusively activated by designer drugs (KORD) (Vardy et al, 2015) into the mPFC. We then conducted single-unit recordings in the mPFC (Figures 2.2a and 2.2b) in behaving mice (Figure 2.2c and Figure 2.3). After subcutaneously injecting the mice with the KORD agonist salvinorin B (SalB), we observed neurons that decreased the amplitude of both their baseline firing and task-related activity (Figures 2.2d,e, and 2.3b,c). Overall, more than 60% of the recorded neurons suppressed their averaged firing rates (baseline and task-related) to below half of their pre-injection rates (Figure 2.2f-h). On average, these suppressed neurons decreased their firing rates to less than 20% of their pre-injection rates within 15 minutes post-SalB injection (Figure 2.2f). We separately analyzed any remaining task-related activity that these neurons displayed. We computed normalized firing rates during the one second following odor onsets, plotted this over the recording session, and confirmed that task-related activity of these
Figure 2.2 | KORD inactivates mPFC neurons. (a) Schematic showing angle of tetrode implantation into the mPFC, in 2 mice that were also injected with KORD in the mPFC. We recorded from tissue in which we expressed KORD, with the goal of quantifying the time course and efficacy of KORD. (b) Coronal section from one mouse recorded in mPFC, showing KORD expression in green. The hole in the tissue is the electrolytic lesion created at the end of the tetrodes, following completion of the experiment, showing that tetrodes were located in a region that expressed KORD. (c-e) Task and firing patterns for a representative neuron, before and after SalB injection. This neuron showed tonic baseline firing prior to odor onset, firing after both cue and reward, and sustained firing between cue and reward. After SalB administration, the neuron still showed firing to cue and reward. However, the tonic baseline firing rate, as well as the sustained firing between cue and reward, decreased after SalB. (f) Normalized firing rates before and after injection, plotted for all inhibited neurons (neurons were categorized as ‘inhibited’ if they suppressed firing rate to less than half of baseline firing rate). Each blue line is an average of all inhibited neurons recorded in one day. Each neuron’s firing rate was normalized to its average pre-injection firing rate. Inset pie charts denote the number of neurons categorized as ‘inhibited’ in the SalB and Saline + DMSO conditions. (g) Baseline firing rate versus post-injection (>15 minutes following SalB or Saline injection) firing rate, on a log-log scale. Points below the unity line show neurons whose firing rates were lower after drug (or saline) injection than before injection. (h) Post-injection firing rate as a proportion of baseline firing rate, for each neuron. Neurons are rank-ordered from most suppressed to most excited. Points below the horizontal line show neurons whose firing rates after drug (or saline) injection were lower than before injection.
neurons displayed a level of suppression similar to that of the averaged recorded activity (Figure 2.3c). Finally, we found that the entire population of recorded neurons contained neurons that were maximally activated at distinct timepoints between odor onset and reward, tiling the entire interstimulus interval when arranged from earliest to latest (Figure 2.3d). Following SalB injection, neural activity no longer spanned the entire interval, instead showing activation immediately following odor onset (Figure 2.3d,e). Taken together, our mPFC recordings confirm that KORD reliably suppressed both baseline and task-related activity, and abolished sustained activity during a classical conditioning task.
Behavior and electrophysiology

We next inactivated mPFC while recording from dopamine neurons in the ventral tegmental area (VTA) (Figures 2.4a, 2.5a). We injected KORD unilaterally into the mPFC of 8 mice (Figure 2.5b). To unambiguously identify dopamine neurons, we expressed channelrhodopsin-2 (ChR2) in dopamine neurons. We classified neurons as dopaminergic if they responded reliably with short latency to pulses of blue light delivered through an optical fiber positioned near our electrodes, which we implanted ipsilateral to the KORD injection (see Materials and Methods, Figures 2.4b-e). On each dopamine recording day, we alternated between subcutaneously injecting animals with saline (control) or SalB (mPFC...
inactivation) at the beginning of the recording session. We trained 4 mice each on Tasks 1 and 2 (Figure 2.4f). Both classical conditioning tasks varied the timing between odor cue and reward (inter-stimulus interval – ‘ISI’). On Odor A trials, the ISI was drawn from a discretized Gaussian distribution ranging from 1.2s to 2.8s, with an average ISI of 2.0s. Odor B and C trials had constant ISIs of 1.2s and 2.8s, respectively. Odor D trials were unrewarded. The only difference between Tasks 1 and 2 was that 100% of Odor A-C trials were rewarded in Task 1, whereas only 90% of Odor A-C trials were rewarded in Task 2. In both tasks, animals learned to lick in anticipation of reward in Odor A-C trials (Figure 2.4g, anticipatory licking for Odors A-C in Tasks 1 and 2 ≠ baseline; $F_{1,36} > 32$, $p < 1.9 \times 10^{-6}$ for all comparisons, one-way analysis of variance (ANOVA)), but not in Odor D trials ($F_{1,41} < 1.4$, $P > 0.23$ for all comparisons, one-way ANOVA). We performed several analyses to ask whether behavior differed between Tasks 1 and 2, and whether behavior differed between Saline and SalB conditions (Figure 2.6). Animals

Figure 2.5 | VTA recording sites and KORD expression. (a) Recording sites in Task 1 and Task 2 animals. (b) Raw histology images for two representative mice from Tasks 1 and 2, respectively. (c) While striatal labeling appears bright in (b), 63x confocal images reveal this striatal labeling to consist predominantly of axons, in contrast to cell bodies that are labeled in the mPFC. (d-e) Areas outlined in black denote the site of KORD expression in individual Task 1 and Task 2 animals. Blue areas in (d) indicate minimum area covered in all Task 1 animals, and red areas in (e) indicate minimum area covered in all Task 2 animals. (f) Minimum areas covered in Task 1 and 2 animals, overlaid.
Figure 2.6 | Analysis of animal behavior in Tasks 1 and 2, Saline and SalB conditions. (a) Comparing the number of licks from odor onset until reward was received (only for trials in which reward arrived at 2.8s, to capture licking throughout the entire possible duration of the ISI). The number of licks were significantly different between Task 1 and Task 2 (# anticipatory licks for Odor A in Task 1 ≠ Task 2; $F_{1,76} = 6.7, p = 1.16 \times 10^{-2},$ one-way analysis of variance (ANOVA)), but not between Saline and SalB conditions, within each task (# anticipatory licks for Odor A in Task 1 SalB ≠ Task 1 Saline; $F_{1,36} = 0.16, p = 0.69;$ # anticipatory licks for Odor A in Task 2 SalB ≠ Task 2 Saline; $F_{1,38} = 0.01, p = 0.91,$ one-way analysis of variance (ANOVA)). (b,c) Comparing the lick rate after Odor A onset until reward is received, in 200ms epochs, between Saline and SalB conditions. We performed 1-way ANOVAs comparing the lick rate between the Saline and SalB condition for each epoch (14 total ANOVAs), for each task. Lick rates were not significantly different in any epoch between Saline and SalB conditions (lick rate for any given epoch after Odor A in Task 1 Saline condition ≠ Task 1 SalB condition; $F_{1,36} < 4.0, p > 0.40$ for all comparisons; lick rate for any given epoch after Odor A in Task 2 Saline condition ≠ Task 2 SalB condition; $F_{1,38} < 4.0, p > 0.40$ for all comparisons, one-way analysis of variance (ANOVA) – no correction for multiple comparisons). (d,e) Best-fit lines relating the number of licks to time following Odor A (but before reward onset) were plotted for behavior during individual recording sessions in Tasks 1 and 2. The best-fit lines and slopes (m) for these two example sessions are shown in green. (f) The slopes were significantly different between Task 1 and Task 2 (slopes in Task 1 ≠ Task 2; $F_{1,76} = 6.1, p = 1.6 \times 10^{-2},$ one-way analysis of variance (ANOVA)), but not between Saline and SalB conditions, within each task (# anticipatory licks for Odor A in Task 1 SalB ≠ Task 1 Saline; $F_{1,36} = 0.87, p = 0.36;$ # anticipatory licks for Odor A in Task 2 SalB ≠ Task 2 Saline; $F_{1,38} = 0.04, p = 0.84,$ one-way analysis of variance (ANOVA)).
Figure 2.6 (Continued). (g) Best-fit sigmoid curves relating the lick rate to time following Odor A (but before reward onset) were plotted for behavior during individual recording sessions in Tasks 1 and 2. The best-fit curves for these two example sessions are shown in green. (h) The timepoints of halfway-to-maximum lick rates were significantly different between Task 1 and Task 2 ($x_{50}$ in Task 1 ≠ Task 2; $F_{1,76} = 7.69$, $p = 7.0 \times 10^{-3}$, one-way analysis of variance (ANOVA)), but not between Saline and SalB conditions, within each task ($x_{50}$ in Task 1 SalB ≠ Task 1 Saline; $F_{1,36} = 0.02$, $p = 0.90$; $x_{50}$ in Task 2 SalB ≠ Task 2 Saline; $F_{1,38} = 0.42$, $p = 0.52$, one-way analysis of variance (ANOVA)). (i) The steepness of the sigmoid fits were not significantly different between Task 1 and Task 2 (steepness in Task 1 ≠ Task 2; $F_{1,76} = 0.59$, $p = 0.44$, one-way analysis of variance (ANOVA)), nor were they different between Saline and SalB conditions, within each task (steepness in Task 1 SalB ≠ Task 1 Saline; $F_{1,36} = 0.3$, $p = 0.59$; steepness in Task 2 SalB ≠ Task 2 Saline; $F_{1,38} = 0.05$, $p = 0.82$, one-way analysis of variance (ANOVA)). (j) Examples of the sigmoid fits for Odor B and C trials, from one example session. The Odor C licking trace is in orange, and the Odor B licking trace is in green. (k) The time of reward ($t_{\text{reward}}$) minus the time to half the maximum lick rate ($x_{50}$), for every individual session. We used this as a measure of temporal imprecision: the smaller $x_{50} - t_{\text{reward}}$, the more precisely the animal licked exactly when water was delivered. (l) Based on the behavioral imprecision computed in (k), we attempted to predict the increase in dopamine RPEs from Odor B to Odor C rewards. However, this analysis did not reveal a significant correlation. For this reason, we chose the Weber fraction in our model based on other animal timing studies, rather than behavior from our tasks.

licked more (# anticipatory licks for Odor A in Task 1 ≠ Task 2; $F_{1,77} = 6.7$, $p = 1.2 \times 10^{-2}$, one-way analysis of variance (ANOVA); Figure 2.6a), and ramped up their lick rates sooner (timepoint halfway to maximum lick rate in Task 1 ≠ Task 2; $F_{1,77} = 7.7$, $p = 7.0 \times 10^{-3}$, one-way analysis of variance (ANOVA)), in Task 1 compared to Task 2 (Figure 2.4g, Figures 2.6g,h). SalB did not affect the number of licks (# anticipatory licks for Odor A in Task 1 SalB ≠ Task 1 Saline; $F_{1,37} = 0.16$, $p = 0.69$; # anticipatory licks for Odor A in Task 2 SalB ≠ Task 2 Saline; $F_{1,39} = 0.01$, $p = 0.91$, one-way analysis of variance (ANOVA), Figure 2.6a), nor did SalB affect multiple other measures that quantified the pattern of licking across time (Figure 2.6b-i).

Finally, prior to analyzing dopamine responses in more depth, we asked whether the types of VTA neurons recruited to the task differed between Saline and SalB conditions. We applied k-means clustering to all of our recorded VTA neurons (counted as those within 500um of an
optogenetically-identified dopamine neuron; n = 761 neurons) and sorted neurons in three clusters that showed phasic activity to cue and reward, sustained positive activity, and sustained negative activity (Figure 2.7a) (Cohen et al., 2012; Eshel et al., 2015; Tian and Uchida, 2015). Based on this analysis, we did not find appreciable differences in the types of activity patterns of recorded VTA neurons, between Saline and SalB conditions in each task (Figure 2.7b).

mPFC inactivation impaired dopamine responses in Task 2, but not in Task 1

We next analyzed averaged dopamine activity on control (saline injection) and inactivation (SalB injection) days. In control sessions, we found that averaged reward responses following Odor A showed opposite trends of temporal modulation between Tasks 1 and 2, replicating our earlier results (Starkweather et al., 2017). That is, in Task 1, post-reward responses decreased as a function of time (Figure 2.8a, colored lines; $F_{8,328} = 12.6$, $p = 9.1 \times 10^{-16}$, 2-way ANOVA; factors: ISI, neuron). In contrast, in Task 2, post-reward responses increased as a function of time (Figure 2.8b, colored lines; $F_{8,320} = 3.7$, $p = 3.8 \times 10^{-4}$, 2-way ANOVA; factors: ISI, neuron). Importantly, scalar timing uncertainty could not account for the positive
Figure 2.8 | mPFC inactivation impaired temporal modulation of dopamine reward responses in a non-deterministically rewarded task (Task 2) but not in a deterministically rewarded task (Task 1). (a) Average non-normalized PSTH for all 42 dopamine neurons recorded during Odor A trials in Task 1, Saline condition. Colored lines are post-reward responses at various timings, and black line is firing prior to reward. Both post-reward (50-200ms following reward) and pre-reward firing (0-200ms prior to reward) are significantly modulated by time. Post-reward firing (mean +/- SEM shown in plots): $F_{8,328} = 13, p = 9.1 \times 10^{-16}$, 2-way ANOVA; factors: ISI, neuron; Pre-reward firing (mean +/- SEM shown in plots): $F_{8,328} = 15, p = 3.5 \times 10^{-11}$. Green and orange dots in insets denote post-reward firing to rewards delivered in Odor B (green) and Odor C (orange) trials, which have constant ISIs of 1.2s and 2.8s, respectively (see Figure 2.4f). (b) Average PSTH for all 41 dopamine neurons recorded during Odor A trials in Task 2, Saline condition. Post-reward firing: $F_{8,320} = 3.7, p = 3.8 \times 10^{-4}$; pre-reward firing: $F_{8,320} = 14, p = 5.3 \times 10^{-18}$. Positive temporal modulation cannot be explained by ISI length alone, as there was a significant difference between the post-reward response for the latest possible Odor A reward delivery, and Odor C reward delivery, which both had ISIs of 2.8s ($Odor\ A_{ISI=2.8s} \neq Odor\ C\ response, F_{1,41} = 22.4, p = 2.7 \times 10^{-5}$, 2-way ANOVA; factors: ISI, neuron). (c) PSTH for all 42 dopamine neurons recorded during Odor A trials in Task 1, SalB condition. Post-reward firing: $F_{8,328} = 11, p = 3.3 \times 10^{-14}$, 2-way ANOVA; factors: ISI, neuron; Pre-reward firing: $F_{8,328}=16, p = 2.4 \times 10^{-19}$. (d) Average PSTH for all 47 dopamine neurons recorded during Odor A trials in Task 2, SalB condition. Post-reward firing: $F_{8,368} = 0.69, p = 0.70$; pre-reward firing: $F_{8,368} = 14, p = 2.7 \times 10^{-17}$. (e) Bar graph showing average slope relating the ISI to RPE magnitude (mean +/- SEM), for all recorded neurons in each condition. The distribution was significantly different in Task 2 ($median_{Saline} \neq median_{SalB}, z = 2.1, p = 3.8 \times 10^{-2}$, Wilcoxon rank-sum test) but not in Task 1 ($median_{Saline} \neq median_{SalB}, z = 0.14, p = 0.89$, Wilcoxon rank-sum test).
temporal modulation of post-reward responses for Odor A. The post-reward response to Odor C (with an ISI of 2.8s, see Figure 2.4f), was significantly smaller than the post-reward response to the latest possible Odor A reward, which also had an ISI of 2.8s (Figure 2.8b; Odor A\textsubscript{ISI=2.8s} ≠ Odor C reward response, \(F_{1,41} = 22.4, p = 2.7 \times 10^{-5}\), 2-way ANOVA; factors: ISI, neuron). In addition, pre-reward firing rates decreased over time (Figures 2.8a,b, black lines; \(F_{8,328} > 9.0, p < 3.5 \times 10^{-11}\) for both groups, 2-way ANOVA; factors: ISI, neuron) in both Tasks 1 and 2.

In inactivation sessions, post-reward dopamine responses in Task 1 remained similar to the control data: post-reward responses decreased as a function of time (Figure 2.8c, colored lines; \(F_{8,328} = 11, p = 3.3 \times 10^{-14}\), 2-way ANOVA; factors: ISI, neuron). Strikingly, in Task 2, mPFC inactivation abolished the pattern of increasing post-reward RPEs across time (Figure 2.8d, colored lines; \(F_{8,368} = 0.69, p = 0.70\), 2-way ANOVA; factors: ISI, neuron) although the observed responses were overall still smaller (~60%) compared to their responses to unexpected reward (Figure 2.9a,b). Interestingly, pre-reward RPEs in both Tasks 1 and 2 decreased over time, similar to the control condition (Figures 2.8c,d, black lines; \(F_{8,368} > 14, p < 2.7 \times 10^{-17}\) for both groups, 2-way ANOVA; factors: ISI, neuron). To confirm that the selective effect on temporal modulation in Task 2 was not simply due to lack of KORD efficacy in Task 1, we simultaneously recorded mPFC neurons in one Task 1 animal (Figures 2.10a,b). We confirmed that KORD inhibited the majority of mPFC neurons in this Task 1 animal (Figures 2.10c-e). We also confirmed that animals injected with SalB, which did not express KORD, displayed intact temporal modulation of post-reward responses in Task 2, confirming that mPFC inactivation, rather than SalB itself, accounted for the observed effects (Figure 2.10f-h). In summary, mPFC
inactivation selectively abolished temporal modulation of post-reward responses in Task 2, in which reward was delivered probabilistically (Figure 2.8e).

To directly ask whether SalB and task identity modulated the effect of reward timing on dopamine responses, we performed an ANOVA that included the following factors: time of reward delivery (t\text{reward}) \times Task 1, t\text{reward} \times Task 2, t\text{reward} \times Drug \times Task 1, t\text{reward} \times Drug \times Task 2, and free water response. ‘Task 1’, ‘Task 2’, and ‘drug’ had values of 1 or 0 depending on the task, and whether or not drug was present. Free water response was included as a factor because dopamine neurons show great variability in the magnitude of reward responses (Eshel et al., 2015), which could not be accounted for by including ‘neuron’ as a variable because different sets of neurons were recorded between conditions. t\text{reward}
Task 1, t_reward \times Task 2, and t_reward \times drug \times Task 2 explained a significant level of variance in the post-reward responses (t_reward \times Task 1: F_{8,1347} = 2.2, p = 2.7 \times 10^{-2}; t_reward \times Task 2: F_{8,1347} = 3.2, p = 1.1 \times 10^{-3}; t_reward \times drug \times Task 2: F_{9, 1347} = 2.6, p = 6.2 \times 10^{-3}), while t_reward \times drug \times Task 1 did not explain a significant level of variance in the post-reward responses (t_reward \times drug \times Task 1: F_{9, 1347} = 0.79, p = 0.62). Therefore, Task and Drug, only in the case of Task 2, interacted
with the timing variable to explain a significant proportion of variance in post-reward responses. This supports our observation that dopamine RPEs show distinct patterns of modulation over time between Tasks 1 and 2, and that SalB only affects this pattern of temporal modulation in Task 2.

Individual neurons recorded in Task 2 tended to show negative temporal modulation, following mPFC inactivation.

Next, we asked how the responses of individual neurons were affected by mPFC inactivation in Task 2. For individual neurons’ responses, we plotted a best-fit line relating the ISI to the number of spikes fired in response to reward on every Odor A trial (Figure 2.11a-d). In the control condition, 29 out of 41 (70%) of neurons displayed positive slopes ($m > 0$) while 12 out of 41 (29%) displayed negative slopes (Figure 2.11f). In contrast, 24 out of 47 (51%) neurons recorded in Task 2 were more likely to show negative modulation following mPFC inactivation. (a-d) PSTHs and raster plots for single neurons recorded in all conditions. A best-fit line was drawn through a plot relating the ISI to the post-reward firing rate for each odor A trial in Task 2. The neuron shown in the Task 2 SalB condition (d) showed negative temporal modulation, similar to neurons recorded in Task 1 (a). (e,f) Slopes of best-fit lines in Saline condition (black) and SalB condition (red) for all dopamine neurons recorded in Task 1 (e) and Task 2 (f). Shading indicates $p < 0.05$, or a 95% confidence interval for the slope coefficient that excludes 0. The distributions were different in Task 2 ($\text{median}_{\text{Saline}} \neq \text{median}_{\text{SalB}}, z = 2.1, p = 0.04$, Wilcoxon rank-sum test) but not in Task 1 ($\text{median}_{\text{Saline}} \neq \text{median}_{\text{SalB}}, z = 0.14, p = 0.89$, Wilcoxon rank-sum test). Inset: For individual animals recorded in Task 1 ($n = 4$) and Task 2 ($n = 4$), variance of slopes in Saline versus SalB condition. The variance increased in Task 2, in the SalB condition ($\sigma^2_{\text{Saline}} \neq \sigma^2_{\text{SalB}}, F_{40, 46} = 0.51, p = 0.03$, $F$-test for equality of two variances), but not in Task 1 ($\sigma^2_{\text{Saline}} \neq \sigma^2_{\text{SalB}}, F_{41, 41} = 0.59, p = 0.10$, $F$-test for equality of two variances).
displayed negative slopes in the inactivation condition (Figure 2.11f). We directly compared the proportion of neurons that showed significant negative or positive slopes, using a Fisher exact test. In Task 1, the proportion of significant slopes showing negative versus positive modulation was not significantly different between drug and control conditions (Figure 2.11e; Saline - 16 negative: 0 positive; SalB – 16 negative, 2 positive, p = 0.49), whereas in Task 2, the proportions significant slopes showing negative versus positive modulation was significantly altered (Figure 2.11f; Saline - 1 negative: 9 positive; SalB – 9 negative: 7 positive, p = 4.1 × 10^{-2}). Therefore, neurons were more likely to show significant negative temporal modulation in Task 2, in the SalB condition.

Accordingly, the distribution of the slopes in the inactivation condition shifted significantly towards smaller values in Task 2 (Figure 2.8e, Figure 2.11f; median_{Saline} ≠ median_{SalB}, z = 2.1, p = 3.8 × 10^{-2}, Wilcoxon rank-sum test). The distribution of slopes was not different in the SalB condition in Task 1 (Figure 2.8e; median_{Saline} ≠ median_{SalB}, z = 0.14, p = 0.89, Wilcoxon rank-sum test). To assess the effect of task and drug on the slopes of all recorded neurons, we performed an ANOVA that included the following factors: Task, Task 1 × Drug, and Task 2 × Drug. In order to compare neurons across groups, we normalized post-reward responses to free water responses prior to computing slopes. This normalization also eased significant deviation from the normal distribution in individual groups, allowing us to perform a valid ANOVA (deviance from normality determined by chi-square goodness of fit test). Task, and Task 2 × Drug, both explained a significant level of variance in the slopes (Task: F_{1,163} = 2.2, p = 1.7 × 10^{-5}; Task 2 × Drug: F_{1,163} = 4.7, p = 3.2 × 10^{-2}), while Task 1 × Drug did not explain a significant level of variance in the slopes (Task 1 × Drug: F_{1,163} = 6.5 × 10^{-2}, p
0.94). Therefore, the distribution of slopes differed over tasks, and SalB only affected the distribution in the case of Task 2.

Finally, because Task 2 slopes tended towards smaller values in the SalB condition, and because more (but not all) Task 2 slopes had negative values, we asked whether the variance of slopes changed significantly in the SalB condition. The variance of the slopes was significantly greater in the inactivation condition, in Task 2 (Figures 2.11f; $\sigma_{\text{Saline}}^2 \neq \sigma_{\text{SalB}}^2$, $F_{40,46} = 0.51$, $p = 0.03$, $F$-test for equality of two variances), with increased variance in the inactivation versus the control condition being reflected in individual animals (Figure 2.11f, black dots). The variance of the slopes was not significantly greater in the inactivation condition, in Task 1 (Figures 2.11e; $\sigma_{\text{Saline}}^2 \neq \sigma_{\text{SalB}}^2$, $F_{41,41} = 0.59$, $p = 0.10$, $F$-test for equality of two variances). Therefore, the abolished temporal modulation shown in Figure 2.8d is the result of averaging the responses of individual neurons that show highly variable trends of temporal modulation, ranging from very negative to very positive modulation across time. Finally, neither the medians nor the variance of post-reward slopes in Task 1 were significantly different between the control and inactivation conditions, further supporting our conclusion that mPFC inactivation selectively affected temporal modulation of post-reward dopamine responses in Task 2.

Our analysis of individual neurons revealed that mPFC inactivation did not simply flatten the response profiles of individual neurons, which could be compatible with the neurons losing the ability to track time. Instead, mPFC inactivation shifted the predominant pattern of positive temporal modulation in Task 2 towards more negative values, with many individual neurons displaying significant negative modulation across time. In other words, many Task 2 dopamine
neurons showed negative temporal modulation similar to Task 1, suggesting that these Task 2 neurons operated as if in a deterministic task regime rather than altogether losing the ability to track time.

**mPFC inactivation spared timing-related aspects of dopaminergic signaling**

To confirm that mPFC inactivation spared time estimation, we analyzed post-reward and reward omission responses for Odors B and C, which had constant ISIs of 1.2s and 2.8s, respectively. Similar to other published accounts, which showed that post-reward RPEs are greater for longer ISIs likely due to scalar timing noise (Fiorillo, Newsome, and Schultz, 2008; Jo and Mizumori, 2015; Kobayashi and Schultz, 2008), we found that post-reward RPEs following Odor C were larger than those following Odor B (Figures 2.12a,b). However, this difference was slight (<1Hz) and was not significantly larger in the inactivation condition ($z = 0.30, p = 0.76$, Wilcoxon rank-sum test), suggesting that temporal uncertainty did not accrue substantially larger during mPFC inactivation. On reward omission
trials, dopamine neurons briefly paused their tonic firing rates at the time of expected reward (Figures 2.12c,d, 2.9d). The decrease in spikes during the omission ‘dip’ was significantly smaller than baseline in both the control and inactivation conditions ($F_{1,47} > 12.4$, $p < 1.3 \times 10^{-3}$ for all groups, 2-way ANOVA; factors: time window, neuron), suggesting that a representation of when reward usually occurs remains intact during mPFC inactivation. This result is consistent with a previous study, which showed that the reward omission response remains intact following mPFC inactivation (Jo and Mizumori, 2015). In summary, mPFC inactivation selectively impaired positive temporal modulation of post-reward responses in Task 2, but spared many other aspects of dopamine responses that required time estimation, including 1) negative temporal modulation of post-reward responses in Task 1, 2) decreasing pre-reward responses in both Tasks, 3) negligible increase in post-reward responses on Odor C versus Odor B trials and 4) precise timing of reward omission ‘dips’ on constant ISI trials.

**Computational modeling implicates mPFC in computing the belief state**

We asked whether we could recapitulate the effect of mPFC inactivation through computational modeling. Dopaminergic RPEs are thought to signal the error term in the temporal difference (TD) learning algorithm (Schultz et al, 1997). The goal of TD learning is to accurately estimate value, defined as the expected discounted cumulative future reward, which is typically approximated as a weighted combination of stimulus features. TD learning uses (putatively dopaminergic) RPEs to update the weights (Sutton 1988). Classically, TD learning utilizes temporal features that track time relative to sensory cues. More recent applications of TD learning to the dopamine system have incorporated hidden state inference by deriving the features from a ‘belief state’, or probability distribution over states (Daw et al, 2006; Rao, 2010;
Figure 2.13 | Computational modeling implicates mPFC in computing the belief state. (a) We modeled Tasks 1 and 2 as Markov decision processes, where each ISI sub-state accounts for 200ms of elapsed time within a trial. The sub-states are partitioned into ISI sub-states and ITI sub-states. In the ISI state, reward will arrive prior to another cue onset; in the ITI state, reward will not arrive prior to another cue onset. Task 1 is a fully observable process because the cue is a completely reliable indicator of the ISI state, hence the 100% probability of entering the first ISI sub-state upon detecting cue onset. Task 2 is a partially observable process because the cue is an unreliable indicator of the ISI state, hence the 10% probability of undergoing a hidden state transition from the ITI back into the ITI, upon detect cue onset. (b) Evolution of the belief state over time. A belief state represents a probability distribution over sub-states (indexed by various colors, as shown in (a)), updated based on sensory information using Bayes’ rule. In the temporal difference (TD) learning model, the sub-state probabilities are linearly combined to produce an estimated value $\hat{V}(t)$. $\delta(t)$ is the reward prediction error used to update the weights $w_i(t)$, where $i$ indexes sub-states. In Task 1, the belief state is fixed with 100% serially occupying the ISI sub-states; in Task 2, the belief state allots more probability to the unrewarded ITI state over time, yielding to the possibility of a reward omission trial. (c, f, i) Averaged PSTHs for 50 simulations of belief state TD model for Tasks 1 and 2. Both the intact model (c), and the model with state inference impaired (f), still display negative temporal modulation of post-reward responses in Task 1, while the timing-impaired model (i) does not. The positive temporal modulation of post-reward responses seen in the intact model of Task 2 (c) becomes blunted upon impairing hidden state inference (f).
Figure 2.13 (Continued). (d,g,j) Same analysis as shown in Figure 2.11, indicating temporal modulation of post-reward RPEs in Task 2, but for simulation outputs. Note that only the manipulation of hidden state inference (g) produces a distribution of post-reward RPEs with temporal modulation that tends more towards negative values, similar to our data (Figure 2.11f). (e,h,k) Averaged PSTHs for 50 simulations of belief state TD model for reward omission responses following Odors B and C. Both the intact model (c), and the model with state inference impaired (f), still display reward omission responses at the time of expected reward, while the timing-impaired model (i) does not.

Starkweather et al, 2017; Lak et al, 2017), which in our tasks reflects the probabilities of the ISI (‘reward will come’) and ITI (‘reward will not come’) states. We modeled Tasks 1 and 2 as Markov decision processes, with the ISI and ITI states comprising sub-states (Figure 2.13a). We chose to explicitly model the ISI and the ITI because these two states dictate whether reward is expected or not. Each sub-state corresponds to a discrete amount of time during the task. Because the Gaussian interval during which reward could be received was discretized into 200ms bins, we modeled each sub-state as corresponding to 200ms. The ISI state comprised 14 sub-states because the longest possible ISI was 2.8s. The ITI state comprised just 1 sub-state because the ITI was drawn from an exponential distribution (to ensure a flat hazard rate). Therefore, the dwell time in the ITI could be modeled with one 200ms sub-state with a high self-transition probability. In Task 1, Odor A onset would correspond to a 100% likelihood of a state transition from the black ITI state (Figure 2.13a, parameters in Figures 2.14a,b) to the first pink ISI substate. As time elapses during the trial, each ISI substate would transition to the subsequent ISI substate, until reward is received and the model transitions back to the black ITI state. Because in Task 1 the cue reliably indicates that the animal is in the ISI state, the model’s belief state is fixed at 100% in the ISI state after the animal observes a cue. In other words, the model allots 100% of its belief, sequentially, into each of the ISI sub-states, as time elapses during the trial (Figure 2.13b, 2.15a,b). In contrast, Task 2 is a hidden Markov decision process because
the cue may lead to an omission trial in 10% of cases (Figure 2.13a, parameters in Figures 2.14a,c). This is modeled as a 10% likelihood of a ‘hidden’ state transition from the ITI state back to the ITI state, without a reward, when a cue is observed. Therefore, which state (ISI versus ITI) the animal is in is not directly signaled. The animal’s actual state is “hidden” because it cannot be reliably deciphered from sensory cues alone. Upon experiencing Odor A onset, the model allots 90% of its belief into the first pink ISI sub-state, and allots 10% of its belief into the black ITI state (Figures 2.13b, 2.15c,d). As time elapses and no reward is received, the model yields to the belief that the current trial is unrewarded, allotting smaller probabilities to each subsequent ISI sub-state, and larger probabilities to the ITI state (Figures 2.13b, 2.15c,d).

The belief state TD model, trained on our tasks, reproduced key aspects of our data. Post-reward firing in Task 1 decreased over time (Figure 2.13c, left panel) because the model learned higher weights for later features, reflecting the higher momentary probability of receiving reward at later timepoints (Figures 2.15a,b).

Figure 2.14 | Computational model details. (a) Observation and Transition matrices for belief state TD model (see Methods). The transition matrix was blurred by a Weber fraction of 0.04 (per sub-state; 0.20 per second). (b,c) The only differences in model values between Tasks 1 and 2 are noted. Most notably, the observation matrix value specifying the likelihood of observing a cue, given the transition back into the ITI, is 0 in Task 1 and non-zero in Task 2. (d) Impaired timing was simulated by blurring the transition matrices by a larger Weber fraction (see Methods). (e) Impaired hidden state inference was simulated by eliminating the non-zero likelihood of observing cue given an ITI→ITI transition, abolishing the model’s ability to acknowledge the possibility of this hidden state transition during an omission trial.

82
Post-reward firing in Task 2 increased over time (Figure 2.13c, right panel), reflecting the model’s mounting belief that it is in an omission trial at later timepoints (Figures 2.15c,d). For individual Task 2 simulations, we computed a best-fit line relating ISI and post-reward RPEs, identical to the slope distribution analysis shown in Figure 2.11. This revealed that the majority of simulations produced positive temporal modulation over time (Figure 2.13d). Pre-reward firing decreased over time in both Tasks 1 and 2 (Figure 2.13c, black lines). Finally, our model produced omission responses around the time of expected reward, for Odors B and C (Figure 2.13e).

To simulate a deficit in hidden state inference, we altered the model parameters such that the model failed to acknowledge the 10% likelihood of an omission trial in Task 2 (Figure 2.14e), while keeping other
Task 2 model parameters—namely the weights that it had learned, and the transition matrix that reflects the temporal structure of the task—intact. In other words, the Task 2 belief state (Figure 2.13b) was fixed with 100% probability in the ISI, and remained uniformly ‘stuck’ at this probability, similar to the Task 1 belief state (Figure 2.15e). Importantly, the probability mass allotted to ISI sub-states still tracked time during each trial by sequentially passing from one sub-state to the next. Rather, the probability assigned to each sub-state was changed by our impairment. While in a model with an intact belief state, the ITI state accrues greater values as time elapses, and accordingly, the sum of the probability allotted to the ISI sub-states decreases over time (Figure 2.15d), the values of these probabilities do not change over time in the impaired model (Figure 2.15e). We found that we could recapitulate our mPFC inactivation data by running 60% of the simulations with this impoverished belief state. Indeed, our inactivation was likely partial, as not all neurons showed inhibition upon mPFC inactivation (Figure 2.2f), and the contralateral (intact) mPFC may communicate with the recorded hemisphere through crossing corticothalamic projections (Vertes, 2001; Vertes, 2004; Gabbott et al, 2005). While averaged responses from Task 1 simulations continued to show post-reward responses that decreased as a function of time, averaged RPEs from Task 2 simulations were flattened across time (Figures 2.13f). Similar to

Figure 2.16 | Cartoon hypothesis of mPFC inactivation results, based on computational modeling. (a) Intact and impaired Task 2 belief states lead to the patterns of RPEs shown in (B), respectively. (b) Differing patterns of RPEs in dopaminergic neurons unaffected and affected by the mPFC inactivation, respectively, according to our hypothesis that mPFC inactivation impaired the belief state in a subset of recorded neurons. (c) Averaging these two patterns together results in flattened averaged RPE, similar to the blunted pattern of temporal modulation seen in Figure 2.8d.
our mPFC inactivation data, Task 2 pre-reward RPEs continued to decrease as a function of time. Furthermore, a greater proportion of Task 2 simulations showed negative temporal modulation of post-reward RPEs than in the intact model (Figure 2.13g, compare with intact model in Figure 2.13d), similar to our inactivation data (Figure 2.11f). These negatively modulated post-reward RPEs occurred in simulations with the corrupted belief state (Figures 2.15e and 2.16). Finally, simulations run with a deficit in hidden state inference still exhibited a reward omission response around the time of expected reward (Figure 2.13h). An omission response occurred despite the belief state being fixed, because later sub-states accrued very low weights (see weights in Figures 2.15c,d), forcing estimated value to drop as soon as the model experienced later time points. Therefore, impairing hidden state inference recapitulated our data: while flattening post-reward responses in Task 2, all other aspects of dopamine signaling were spared.

We next examined the effect of impairing the model’s timing mechanism by blurring the transition probabilities between sub-states (Takahashi et al, 2016; see Materials and Methods, Figure 2.14d). We found that this timing impairment in the model could somewhat flatten temporal modulation of Task 2 post-reward responses (Figure 2.13i, right panel). However, blurring the model’s timing estimation affected many other aspects of the simulations, which were inconsistent with our data. For example, Task 1 post-reward responses were also flattened (Figure 2.13i, left panel), pre-reward responses in both Tasks 1 and 2 were flattened (Figure 2.13i, black lines), most simulations still showed positive modulation of post-reward responses over time (Figure 2.13j), and reward omission responses were abolished (Figure 2.13k).
We also attempted to blur the model’s timing estimation less dramatically than shown in Figure 2.14d, but found that this actually increased the positive temporal modulation of post-reward responses in Task 2. This occurred because a smaller increase in scalar timing uncertainty makes only later rewards harder to predict, thereby increasing positive temporal modulation. In contrast, the parameter that we used to blur the transition matrix was so large that both early and late rewards were harder to predict, blunting overall temporal modulation. Finally, we also attempted to simulate a deficit in timing by uniformly blurring the transition matrix (rather than increasing scalar uncertainty). Similar to the previous manipulation, however, this also eliminated reward omission responses, unlike what we observed in our data. Therefore, our data are most consistent with mPFC shaping hidden state inference, rather than timing estimation, in the dopaminergic circuitry.
DISCUSSION

Taken together, our results demonstrate that the dependence of dopamine RPEs on state uncertainty can be explained by a computational framework that incorporates a belief state. Our results further demonstrate that hidden state inference contributing to the dopamine RPE computation critically depends on the integrity of mPFC functioning.

Although past studies have implicated the mPFC in interval timing (Kim et al, 2009; Kim et al, 2013; Xu et al, 2014), the behavioral tasks used in these studies were not designed to disentangle the involvement of mPFC in time estimation itself versus in an inferential process that evolves across time. For instance, Kim and colleagues inactivated the mPFC in a task that required rats to categorize an interstimulus interval as ‘short’ versus ‘long’ (Kim et al, 2009). Rats’ psychometric curves (correctly categorizing ‘short’ versus ‘long’) became flattened as a function of time, post-mPFC inactivation. This task can be conceptualized as requiring hidden state inference, in addition to interval timing. That is, based on the time interval, the animal infers the correct (hidden) categorization. As time elapses during the interstimulus interval, the animals’ belief state must increasingly favor the ‘long’ category. Blunting this dynamic modulation of the belief state, rather than impaired time estimation per se, could explain that phenotype observed upon mPFC inactivation. Furthermore, another study showed that hidden state inference plays an important role in a simple sensory discrimination task using ambiguous visual stimuli (Lak et al, 2017). As illustrated by these examples, hidden state inference may underlie diverse neural processes, such as timing and sensory discrimination, making it difficult to understand the contribution of brain regions to one process in particular. In the present study,
we experimentally separated hidden state inference and timing by assaying the contribution of
the mPFC through two different behavioral tasks.

Our results demonstrated the necessity of mPFC in hidden state inference, but not in
interval timing, suggesting that these two processes have separable neural substrates. In our task,
inferring the hidden state required the animal to know how much time had passed. For this
reason, timing information must be routed from elsewhere to the mPFC, where the belief state is
computed. Where, then, is interval timing computed? Two different lesion studies have shown
that lesioning the ventral striatum (Takahashi et al, 2016) or lateral habenula (Tian and Uchida,
2015) resulted in dopamine neurons losing their ability to ‘pause’ at the time of an unexpected
reward omission. Furthermore, another study showed that individual neurons in the striatum
show bursts of activity that span the ISI of a lever-pressing task, and re-scale the absolute time of
their bursting activity to tile longer ISIs (Mello et al, 2015). Therefore, the striatum contains
neurons that flexibly represent behaviorally relevant time intervals, and could convey timing
information to the dopamine system through a pathway involving the lateral habenula.

How could belief state information be conveyed from the mPFC to dopamine neurons?
Several possible routes exist. First, the mPFC sends predominantly ipsilateral projections directly
to VTA dopamine neurons (Carr and Sesack, 2000; Vertes, 2004; Gabbott et al, 2005; Watabe-
Uchida, 2012), providing a direct route by which mPFC activity could influence dopamine
signaling. Other routes involve multiple synapses and potential relays. For example, the mPFC
sends dense ipsilateral projections to the striatum, including the nucleus accumbens (Sesack,
Deutch, Roth, and Bunney, 1989; Vertes, 2004; Gabbott et al, 2005), which then supplies major
inputs to dopamine neurons in the VTA (Watabe-Uchida, 2012). Consistent with abundant mPFC projections to striatum, we observed bright striatal labeling in animals injected with KORD. Another route is through the mediodorsal thalamus, which receives both ipsilateral and contralateral input from the mPFC (Vertes, 2001; Vertes, 2004; Gabbott et al, 2005). This thalamic nucleus is richly interconnected with the mPFC and other prefrontal cortical regions (Mitchell, 2015), and is a potentially important node in sustaining persistent activity in the mPFC by analogy to other recurrent corticothalamic circuits (Guo et al, 2017). Furthermore, this corticothalamic pathway could provide a route by which some intact belief state information reaches dopamine neurons on the recorded hemisphere, accounting for why we observed only a partial effect of inactivation (just 60% of our simulations were run with an impaired belief state to fit our data).

A previous study by Takahashi et al showed that the inactivating the orbitofrontal cortex (OFC) impaired the state representation necessary for computing dopamine RPEs (Takahashi et al, 2011). Furthermore, a previous study using Pavlovian instrumental transfer demonstrated that OFC inactivation is compatible with a deficient version of task state space (Bradfield et al, 2011). In our experiment, we avoided expressing KORD in both the medial and lateral OFC (Figure 2.5). Our mPFC inactivation produced a different, but related, impairment. Rather than ablating the state representation itself, mPFC inactivation impaired the brain’s ability to dynamically infer states only when they were hidden, freezing the inferred probability distribution over states that should have evolved over time. Here we argue that a basic representation of observable state space remains intact upon mPFC inactivation. If the state representation were abolished all together, this would mean that the brain could no longer distinguish the ISI from the ITI, even immediately following observable cues such as reward.
Prediction errors in the absence of a hidden state representation would resemble RPEs produced by the complete serial compound representation used in the classic TD model (Schultz et al. 1997). This TD model would not have a ‘reset’ mechanism that switches from the ISI into the ITI as a reward is observed (Suri and Schultz, 1998). The RPEs produced by this TD model would simply reflect the temporal distribution of rewards, effectively reproducing a Gaussian distribution in both Tasks 1 and 2 (see Starkweather et al, 2017, for details). For this reason, our results are incompatible with an ablated state representation. Instead, we conjecture that OFC conveys a state representation to the mPFC (Wilson et al., 2014). This OFC state representation would contain a vector of possible states, the size of which depends on the complexity of the task; in our simple task it could contain just a ‘reward state’ and a ‘no reward state’. The mPFC then computes a probability distribution over these possible hidden states furnished by the OFC. In a more complex example, a task could contain more states (ex. ‘large reward state’, ‘small reward state’, and ‘no reward state’), leaving the mPFC to compute a probability distribution over a greater number of possible states.

Another remaining question is how a state representation, such as the representation of the ISI and the ITI in our task, could be formed in the first place. Computational hypotheses that address this question exist (Gershman, Norman, and Niv, 2015), although fewer experiments have attempted to link these hypotheses to the dopamine system. In our tasks, we purposefully made the time between trials highly unpredictable—the inter-trial interval was drawn from an exponential distribution, making it impossible for the animal to predict when a cue would come on. However, once a cue does come on, it has high ‘temporal informativeness’ (Balsam and Gallistel, 2009), because it reliably predicts an upcoming appetitive reward at a somewhat
predictable timing. The cue is the one piece of information the animal can rely on, to provide information about when a reward would arrive. Therefore, it is natural to conceive of the task in two pieces—the nebulous state before cue onset, in which the animal cannot know when to expect observable stimuli; and the state after cue onset, in which the animal knows that a reward is likely to come. One idea is that the animal is able to identify temporally informative stimuli in the environment, and forms a state representation based on this. A candidate region that might be attuned to detecting these coincidences is the hippocampus, which has been theorized to generate a predictive map that respects the transition structure within a (usually spatial) task (Stachenfeld, Botvinick, and Gershman, 2017). Such a predictive map, if applied instead to transitions in time rather than space, is exactly what is needed to build a useful state representation in a classical conditioning task with informative temporal, not spatial, cues. Recently, several studies have suggested that the hippocampus stores temporal relationships between stimuli, in addition to spatial relationships (Deuker et al, 2016; Eichenbaum, 2015; Howard and Eichenbaum, 2014; Oprisan et al, 2018). Thus, the hippocampus could serve as a predictive map that codes expected future occupancies, including temporal relationships, between environmental stimuli. Based on this predictive map, task states useful for reinforcement learning could be formed, potentially in the OFC (Wilson et al, 2014), by emphasizing stimuli closely linked to rewarding stimuli. This state representation from the OFC could then be relayed to mPFC to compute the belief state.

Our inactivation result provides further experimental evidence for the belief state TD model. We previously noted that the belief state TD model is not the only explanation of our dopamine recording data under control conditions (intact mPFC). That is, a TD model that includes a state representation that separates the ISI and the ITI but does not explicitly encode
probabilities could also explain the divergent patterns of response between Tasks 1 and 2 (Starkweather et al, 2017). The microstimulus model, which includes scalar temporal uncertainty, as well as this ‘state reset’, is another example of this type of TD model. ‘State reset’ TD models, such as the microstimulus model, are able to reproduce our data because they learn a different set of weights associated with each of the model’s temporal kernels, depending on the task that they are trained on. In Task 1, these weights increase over time, similar to the value function in Task 1 (Figure 2.15b); in Task 2, these weights decrease for later timepoints, similar to the value function in Task 2 (Figure 2.15d). By shutting down the mPFC, we transiently 'switched' neurons from operating as if they were in a probabilistic regime (Task 2) to operating as if they were in a deterministic regime (Task 1). This phenotype is not easily explained by state reset TD models. If the pattern of weights were neutralized by mPFC inactivation, this would blur temporal modulation in both Tasks 1 and 2, as well as disrupt reward omission responses. Rather, our inactivation is parsimoniously explained by a 'hidden state inference' component being selectively and transiently abolished by our experimental manipulation. For this reason, our inactivation result further validates the belief state TD model.

Could our data be explained by mPFC encoding the risk of a trial going unrewarded, or the memory of unrewarded trials? While reducing the risk of reward omission would flatten the belief state, flattening the belief state (as we did) did not eliminate all aspects of our model that were sensitive to risk, such as the smaller magnitude of learned weights in the 90%-rewarded task (compare weights in Figure 2.15a,c). Here, we argue that our data are not compatible with a general loss of risk-related information. In our tasks, risk comes into play not only on Odor A trials, but also on Odors B and C trials, as these trials were also 90%-rewarded. If all risk-related
information were impaired, we would expect for the magnitude of reward responses on Odor B and Odor C trials to change. For instance, if animals no longer perceived Odor B and C trials as risky, expectation should be higher in the mPFC inactivation condition, resulting in more suppressed reward responses. We directly compared the magnitude of post-reward responses on Odor B and Odor C trials by normalizing to free water responses (in order to compare across different groups of neurons). We found that the responses in the 90%-rewarded task were suppressed to ~55% of free water responses, and were not significantly different between Saline and SalB conditions (Figure 2.9c). Furthermore, these Task 2 responses (90%-rewarded) were significantly bigger than in Task 1 (100%-rewarded) in both the Saline and SalB conditions (suppressed to ~45% of free water responses, Figure 2.9c). This supports our point that RPEs are bigger in riskier conditions than in non-risky (100%-rewarded) conditions, regardless of whether the mPFC was inactivated or not. Therefore, broadly eliminating risk-related information (or the memory of event omissions) cannot explain our data, because it is unclear why this information would only come into play in Odor A trials but not in Odor B and Odor C trials. Rather, altering the model’s ability to modulate the belief state across time provided a better fit for our findings.

How might the neural correlates of a belief state be represented by mPFC activity? Our mPFC inactivation spanned multiple subregions of the mPFC, including both prelimbic (PL) and infralimbic (IL) cortices (Figure 2.5). PL and IL have been dichotomized in their roles in reward-seeking and extinction, respectively (Gourley and Taylor, 2016). In order to promote reward-seeking behavior, the animal’s belief state must favor a rewarded state. Conversely, to promote extinction of a formerly reward-predicting cue, the animal’s belief state must favor an unrewarded state. It is possible that the belief state is represented in both PL and IL, with PL neurons signaling belief in the rewarded (ISI) state, and IL neurons signaling belief in the
unrewarded (ITI) state. A testable prediction of this hypothesis is that, following cue onset in the 90% rewarded task, PL neurons show sustained activation that decreases as time within the trial elapses (similar to the colored bars for Task 2 in Figure 2.13b). In contrast, IL neurons should show ramping that increases as time elapses (similar to the black bars for Task 2 in Figure 2.13b). Furthermore, our model predicts that different effects should be observed on dopamine responses in our task, following PL and IL inactivation. If PL were lesioned, thereby impairing beliefs in the rewarded state and favoring the unrewarded state, RPEs in the 90%-rewarded task should become larger and tend towards more positive temporal modulation, due to the model favoring the possibility of a reward omission trial. If IL were lesioned, thereby impairing beliefs in the unrewarded state and favoring the rewarded state, RPEs should become smaller and favor negative temporal modulation. These hypothetical differences between PL and IL could also produce different behavioral phenotypes, upon differential inactivation. For instance, in the behavioral paradigm discussed in Li et al, 2013 (Li and Dudman, 2013), the authors trained mice on an operant task in which rewards were delivered in a Gaussian distribution of ISIs after the mice pulled a lever. In probe trials, rewards were not given. Upon lesioning PL—an area potentially involved in signaling the rewarded state—mice might wait in the reward port for a shorter duration prior to ‘giving up’ on these omission trials and re-initiating a trial, whereas the opposite effect would be expected upon lesioning IL. In future experiments, it would be important to characterize the activities of these various subregions of the mPFC, with the possibility of disentangling the contribution of PL versus IL to representing a probability distribution over favorable and unfavorable hidden states, respectively.
Inference based on ambiguous information is a fundamental computation that the brain must perform in natural environments (Fiser et al, 2010; Pouget et al, 2013). Our findings represent an important conceptual advance in understanding the contributions of mPFC to reinforcement learning under conditions of state uncertainty.
METHODS

Animals. We used 12 adult male mice, heterozygous for the transgene that expresses Cre recombinase under the control of the DAT promoter (B6.SJL-Slc6a3tm1.1(cre)Bkmm/J, The Jackson Laboratory, Backman et al, 2007). 4 animals were used in Task 1, 6 animals were used in Task 2 (2 of these control animals did not express KORD), and 2 animals were used to test KORD (see summary table below). Animals were housed on a 12-h dark/12-h light cycle (dark from 7AM to 7PM). We trained animals 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 and approved by the Harvard Institutional Animal Care and Use Committee.

Surgery and viral injections. We performed all surgeries under aseptic conditions with animals under isoflurane (1-2% at 0.5-1.0L/min) anesthesia. Analgesia (buprenorphine, 0.1mg/kg, intraperitoneal) was administered pre-operatively and at 12-h checkpoints post-operatively. We performed some permutation of 4 different surgeries on each mouse, summarized in the table below. To record dopaminergic neurons in the VTA, we performed two surgeries that stereotactically targeted the left VTA (from bregma: 3.1mm posterior, 0.6mm lateral, 4.2mm ventral). To express ChR2 in the left VTA, we injected 500nL of adeno-associated virus (AAV, serotype 5) carrying an inverted ChR2 (H134R) fused to the fluorescent reporter eYFP and flanked by double loxP sites (Cohen et al, 2012; Atasoy et al, 2008) into the left VTA. We previously showed that the expression of this virus is highly selective and efficient in dopamine neurons (Cohen et al, 2012). After 2 weeks, we performed a second surgery to implant a head plate and custom-built microdrive containing 8 tetrodes and an optical fiber. To express KORD
in the mPFC, we injected 1uL of adeno-associated virus (AAV, serotype 8) carrying KORD fused to the fluorescent reporter mCitrine and downstream of a CaMKIIa promoter (Vardy et al, 2015). We injected ~100nL in 9 different injection sites, with the injection needle angled 22.5 degrees to the normal line, from a coronal view (from bregma: (1) 1.42mm anterior, 0.8mm l, 1.46mm ventral relative to injection angle; (2) 1.42mm a, 1.2mm l, 2.24mm v; (3) 1.70mm a, 0.8mm l, 1.5mm v; (4) 1.70mm a, 1.2mm l, 2.26mm v; (5) 1.98mm a, 0.75mm l, 1.32mm v; (6) 1.98mm a, 1mm l, 2.07mm v; (7) 2.34mm a, 0.8mm l, 1.33mm v; (8) 2.68mm a, 0.67mm l, 0.93mm v; (9) 2.96mm a, 0.6mm l, 0.676mm v). This KORD injection was performed during the same surgery as the ChR2 injection surgery, if the animal was to have both surgeries. If the animal was to be implanted with tetrodes in the mPFC, we implanted tetrodes 3 weeks later. We implanted tetrodes in the mPFC at a 30 degree angle to the normal line, from a sagittal view, in order to sample the anterior-posterior and dorsal-ventral axes as we moved our drive (from bregma: 2.7mm anterior, 0.25mm lateral, 0.6mm ventral relative to injection angle).

**Table 2.1 | Summary of animals used in Chapter 2.**

<table>
<thead>
<tr>
<th></th>
<th>KORD injected into mPFC</th>
<th>ChR2 injected into VTA</th>
<th>Tetrodes implanted into mPFC</th>
<th>Tetrodes implanted into VTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 mice (KORD testing)</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>2 mice (no KORD control, Task 2)</td>
<td></td>
<td>X</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>4 mice (Task 1)</td>
<td>x</td>
<td>X</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>4 mice (Task 2)</td>
<td>x</td>
<td>X</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>
Behavioral paradigm. After 1 week of post-surgical recovery, we water-restricted mice in their cages. Weight was maintained above 85% of pre-restriction body weight. We habituated and briefly head-restrained mice for 2-3 days before training. Odors were delivered to animals with a custom-made olfactometer (Uchida and Mainen, 2003). Each odor was dissolved in mineral oil at 1/10 dilution. 30μL of diluted odor was placed into glass fiber filter-paper, and then diluted with filtered air 1:20 to produce a total 1L/min flow rate. Odors 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 animals. We automatically detected licks by measuring breaks of an infrared beam 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 9 points ([1.2s 1.4s 1.6s 1.8s 2.0s 2.2s 2.4s 2.6s 2.8s]; mean = 2s; SD = 0.5s), prior to reward delivery. For both Tasks 1 and 2, rewarded odor B and odor C trials consisted of 1s odor presentation followed by either 1.2s or 2.8s delay from odor onset, respectively, prior to reward delivery. In both tasks, odor D trials were unrewarded. In Task 1, reward was given in 100% of trials. In Task 2, reward was given in 90% of trials. For all tasks, reward size was kept constant at 3μL. Trial type was drawn pseudorandomly from a scrambled array of trial types, in order to keep the proportion of trial types constant between sessions. The ITI between trials was drawn from an exponential distribution (mean = 12s) in order to ensure a flat hazard function. Animals performed between 150-300 trials per session.
**SalB injection.** To inject the KORD agonist Salvinorin B (SalB), we placed 1mg SalB and 10uL dimethyl sulfoxide (DMSO - Sigma) in a 2mL Eppendorf tube. After vigorously tapping the tube and ensuring that the SalB powder settled into the DMSO, we sonicated the tube for 1 minute (Branson). We removed the tube from the sonicator and vigorously tapped the tube for 5 seconds, and then sonicated the tube for 1 additional minute. We then added 90uL PBS into the tube, and immediately subcutaneously injected the final 100uL mixture into the mouse. For our dopamine recording experiments, we began the recording session 15 minutes following SalB injection.

**Electrophysiology.** We based recording techniques on previous studies (Cohen et al, 2012; Tian et al, 2015; Eshel et al, 2015). We recorded extracellularly from the VTA using a custom-built, screw-driven Microdrive (Sandvik, Palm Coast, Florida) containing 8 tetrodes glued to a 200µm optic fiber (ThorLabs). Tetrodes were glued to the fiber and clipped so that their tips extended 200-500µm 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 9000 Hz and recorded continuously at 32kHz. To extract spike timing, signals were band-pass-filtered between 300 and 6000Hz and sorted offline using MClust-4.3 (A.D. Redish). At the end of each session, the fiber and tetrodes were lowered by 75um to record new units the next day. To be included in the dataset, a neuron had to be well-isolated (L-ratio < 0.05) and recorded within 300um of a light-identified dopamine neuron (see below) to ensure that it was recorded in the VTA. We also histologically verified recording sites by creating electrolytic lesions using 10-15s of 30µA direct current.
To unambiguously identify dopamine neurons, we used ChR2 to observe laser-triggered spikes (Cohen et al, 2012; Lima et al, 2009; Kvitsiani et al, 2013). 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 10 473nm light pulses, each 5ms long, at 1, 5, 10, 20, and 50Hz, with an intensity of 5-20mW/mm² at the tip of the fiber. Spike shape was measured using a broadband signal (0.1-9,000Hz) sampled at 32kHz. To be included in our dataset, neurons had to fulfill 3 criteria (Cohen et al, 2012; Tian et al, 2015; Eshel et al, 2015):

1) Neurons’ spike timing must be significantly modulated by light pulses. We tested this by using the Stimulus-Associated spike Latency Test (SALT, Kvitsiani et al, 2013). We used a significance value of $P < 0.05$, and a time window of 10ms after laser onset.

2) Laser-evoked spikes must be near-identical to spontaneous spikes. This ensured that light-evoked spikes reflect actual spikes instead of photochemical artifacts. All light-identified dopamine neurons had correlation coefficients $> 0.9$.

3) Neurons must have a short latency to spike following laser pulses, and little ($<3$ms) jitter in spike latency. While others have used a latency criteria of 5ms or less (‘short latency’) (Cohen et al, 2012; Tian et al, 2015; Eshel et al, 2015), we found that the high laser intensity required to elicit this short latency spike sometimes created a mismatched waveform, due to 2 neurons near the same tetrode being simultaneously activated. For this reason, we often decreased the laser intensity and elicited a spike.
5-10ms (‘longer latency’) after laser onset. We separately analyzed neurons in both the ‘short latency’ and ‘longer latency’ categories, and found qualitatively similar results in each group. Therefore, we pooled all dopamine neurons with latencies below 10ms in our analyses.

**Data analysis.** We focused our analysis on light-identified dopamine neurons. To measure firing rates, PSTHs were constructed using 1ms bins. Averaged PSTHs shown in figures were smoothed with a box filter ranging from 80ms (phasic RPEs) to 300ms (reward omission RPEs). Average pre-reward firing rates were calculated by counting the number of spikes 0-400ms prior to reward onset. We also attempted using window sizes ranging from 180-600ms, and these produced similar results. Average post-reward firing rates were calculated by counting the number of spikes 50-200ms after reward onset in both Tasks 1 and 2. We calculated where the baseline-subtracted reward response on an Odor A trial rewarded at 2s (the most abundant trial type) is significantly elevated above zero (p < 0.05, not corrected for multiple comparisons). This window of significantly elevated firing rates is from 50 to 200ms after reward onset. Both pre- and post-reward responses were baseline-subtracted, with baseline taken 0-1s prior to odor onset. Reward omission responses were calculated by counting the number of spikes 0-1000ms after the usual reward delivery time. The number of spikes fired following free reward delivery was calculated by counting the number of spikes 0-300ms after reward onset (this window was wider than that used for post-reward responses for predicted rewards, because free water responses persisted for a longer length of time). Furthermore, we observed that well-trained animals would occasionally ignore reward deliveries outside of odor-cued trials, so we measured free water responses only if we recorded four or more licks in the 1 second following free water delivery.
Note that this was not a very high threshold to exceed (the average lick rate during reward receipt was around 7-8 licks/s), and that lowering this threshold to two or three licks per second yielded similar results in ANOVAs and normalized data.

**Statistics.** No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications (Cohen et al, 2012; Tian et al, 2015; Eshel et al, 2015). Data collection and analyses were not performed blind to the conditions of the experiments. Animals were chosen at random for Tasks 1 or 2. All trial types were randomly interleaved within a single recording session. We verified that all groups of data (including both electrophysiology and behavior) 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 2-factor ANOVA, with neuron and ISI as factor. To test whether individual neurons’ RPEs were modulated by ISI length, we fit a line to the data (dopamine RPEs versus ISI) and reported the slope.

**Code Availability.** Code used to implement the computational modeling in this manuscript can be found in a Supplementary Software section and at this GitHub link: https://github.com/cstarkweather

**Immunohistochemistry.** After 4-8 weeks of recording, we injected mice with an overdose of ketamine/medetomidine. Mice were exsanguinated with saline and perfused with 4% paraformaldehyde. We cut brains in 100um coronal sections on a vibrotome and immunostained with antibodies to tyrosine hydroxylase (AB152, 1:1000, Millipore) in order to visualize
dopamine neurons. We additionally stained brain slices with 4,6-diamidino-2-phenylindole (DAPI, Vectashield) to visualize nuclei. We confirmed AAV expression with eYFP (ChR2) or mCitrine (KORD) 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 AAV (see Figure 2.5a), and to verify that the KORD expression was in the mPFC (Figure 2.5b-f).

Computational modeling.

Belief state TD Model: We simulated TD error signaling in our tasks by using a belief state TD model, similar to that proposed by Daw and colleagues (Daw et al, 2006), as well as Rao (2010), and applied to a previous dataset utilizing Tasks 1 and 2 (Starkweather et al, 2017).

To capture the discrete dwell times in our tasks (1s 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 2.8s after odor onset), we coded a Markov equivalent of a Semi-Markov model (see Daw et al, 2006). The Markov process contained 30 total hidden sub-states (Figures 2.14 and 2.15), with each sub-state corresponding to 200ms in the Tasks. Sub-states 1-5 corresponded to the passage of time during the 1s odor presentation; sub-states 6-14 corresponded to the passage of time preceding the 9 possible reward delivery times (Figure 2.13a). Sub-state 30 corresponded to the ITI. If reward was received at the earliest possible time (1.2s), this would correspond to the model proceeding through sub-states 1-6, and then transitioning to sub-state 30. If reward was received at the latest possible time (2.8s), this would correspond to the model proceeding through sub-states 1-14, and then transitioning to sub-state 30. We included extra ISI substates 15-29 to
accommodate temporal blurring. For instance, if 2.8s had elapsed, this would correspond to a probability distribution centered at sub-state 14, and blurred over neighboring sub-states to an extent that depends on the degree of temporal uncertainty.

In our experiments, the hidden sub-state is known to the experimenter, but not to the animal. How should the animal’s belief over sub-states be updated over time? The normative solution is given by Bayes’ rule:

$$b_i(t + 1) \propto p(o(t)|i) \sum_j p(i|j) b_j(t - 1)$$  \hspace{1cm} (1)

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 $$o(t) \in \{\text{cue, reward, null}\}$$ under hypothetical sub-state $$i$$, and $$p(i|j)$$ is the probability of transitioning from sub-state $$j$$ to sub-state $$i$$.

In TD learning, value is defined as the expected discounted cumulative future reward (Sutton, 1988):

$$V(t) = E \left[ \sum_{\tau=t}^{\infty} \gamma^{\tau-t} r(\tau) \right]$$  \hspace{1cm} (2)

where $$E[\cdot]$$ denotes an average over randomness in reward delivery, $$r(\tau)$$ is the reward at time $$\tau$$, and $$\gamma$$ is a discount factor that down-weights future rewards.

The value function estimate is modeled as a linear combination of stimulus features, which in the belief state TD model is the belief state $$b(t)$$.
\[ \hat{V}(t) = \sum_i w_i b_i(t) \]  

(3)

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

\[ \Delta w_i = \alpha b_i(t) \delta(t) \]  

(4)

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

\[ \delta(t) = r(t) + \gamma \hat{V}(t + 1) - \hat{V}(t) \]  

(5)

In the belief state TD model, it is assumed that the animal has learned a state transition distribution, encoded by matrix T (Figure 2.14a). 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 9 timepoints when reward could occur. For example, the hazard rate of receiving reward at 1.2s would correspond to \( T(6,30) \), or the probability of transitioning from sub-state \( 6 \rightarrow 30 \). 1 minus the hazard rate of receiving reward at 1.2s would correspond to \( T(6,7) \), or the probability of transitioning from sub-state \( 6 \rightarrow 7 \). We captured the exponential distribution of dwell-times in the ITI state by setting \( T(30,30) = 64/65 \) and \( T(30,1) = 1/65 \). An exponential distribution with a hazard rate (ITI_hazard) of 1/65 has an average dwell time of 65. 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 (Figures 2.14b,c):

Task 1:

\[ T(30,30) = 1 - \text{ITI}_\text{hazard} \]
T (30,1) = ITI_hazard

Task 2:

T (30, 30) = 1 - ITI_hazard \times 0.9

T (30,1) = ITI_hazard \times 0.9

This difference in T between Task 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 (Figure 2.14a). There were 3 possible observations: null, cue, and reward. The likelihood of a particular observation given that the hidden state underwent a transition from i→j, was captured as follows:

O (i,j,1) = likelihood of observation of ‘null’, given i→j transition
O (i,j,2) = likelihood of observation of ‘cue’, given i→j transition
O (i,j,3) = likelihood of observation of ‘reward’, given i→j transition

In order to switch from sub-state 30 (ITI) to sub-state 1 (first state of ISI), the animal must have an observation of the cue: O (30,1,2) = 1. In order to switch from sub-state 10 (middle of ISI) to sub-state 30 (ITI), the animal must have an observation of reward: O (10, 30,3) = 1. The only difference in O between Task 1 and Task 2 was as follows (Figures 2.14b,c):

Task 1:

O (30, 30,1) = 1 (null observation)

Task 2:

O (30, 30,1) = 1-ITI_hazard*0.1 (null observation)
This difference in O between Task 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.

Simulating temporal uncertainty in the belief state TD model: To simulate a small amount of scalar timing uncertainty, we blurred the transition matrix by a normal distribution, whose width is proportional to the amount of elapsed time:

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

We used a Weber fraction $\phi = 0.04$ (compounded over five 200ms sub-states, this scales to 0.2 per 1 second), which is similar to values use in other animal timing work (Janssen and Shadlen, 2005; Tsunoda and Kakei, 2008). We used this value rather than compute a Weber fraction based on our behavioral data, because our behavioral data could not predict the increase in post-reward dopamine responses between Odor B and C and thus did not provide a clear correlate of temporal uncertainty (Figure 2.6j-l). We also incorporated uncertainty regarding when cue onset was detected, as the animal’s cue detection is affected by variability in the sniff cycle. We did this by jittering the timing of observations themselves by a normal distribution with a width of 2 sub-states (400ms in real intra-trial time). We justified this choice of width based on measurements on mice sniff cycles (358 ± 131 according to Shusterman et al, 2011). For instance, a true ISI of 10 sub-states would ideally be detected as 10 null observations after cue onset, but could occasionally be detected as 9 or 11 null observations. Both of these manipulations—adding scalar timing uncertainty and jittering the observation of cue timing—allowed us to better match our model to the data for two reasons. First, these manipulations increased the magnitude of
RPEs, even when rewards were predicted. Second, reward omission responses became more smeared in time. Both of these changes occurred because the timing of reward could no longer be perfectly predicted.

Training the belief state TD model and conducting simulations: We first trained the belief state TD model on either Task 1 or Task 2, for 100 sessions, consisting of 100 trials each. We used a learning rate of $\alpha = 0.1$ on all sessions. We used a discount factor of $\gamma = 0.93$. We decreased this value from 0.98, which was used in our previous publication (Starkweather et al, 2017), because it allowed us to better fit the Task 1 data. With $\gamma = 0.98$, the temporal modulation in the 100%-rewarded task appears much smaller than it actually was in our data (Figure 2.8a), because the model would learn a value signal that did not deviate substantially from the earliest possible reward to the latest possible reward (with very shallow discounting). However, with $\gamma = 0.93$, the temporal modulation in the 100%-rewarded better matched our data.

The simulation results presented in the text (Figure 2.13) were produced by running the belief state TD model on either Task 1 (100% rewarded) or Task 2 (90% rewarded), for 50 sessions, consisting of 50 trials each. In all simulations, we used a learning rate of $\alpha = 0$ (assuming learning has already asymptoted) and a discount factor of $\gamma = 0.93$. We added Gaussian white noise to the RPE’s generated by the simulations, by using the MATLAB function awgn and a signal-to-noise ratio of 12.

Impairing the belief state: Our intact belief state model captured state uncertainty because the observation of ‘cue’, in Task 2, was an ambiguous indicator of the ISI versus the ITI:
O (30, 30,2) = ITI_hazard*0.1 (cue in a small percentage of cases)

O (30, 30,1) = 1-ITI_hazard*0.1 (null observation)

We impaired the belief state by fixing O(30,30,1) to 1 and O(30,30,2) to nearly 0 (we could not make it exactly 0 because the model would then be unable to proceed through a new trial following an omission trial). This impairment had the effect of flattening the belief state (Figures 2.16, 2.15e)

**Impairing timing:** We impaired timing by increasing the Weber fraction used to blur the transition matrix. The simulations shown in Figure 2.13 used a range of Weber fractions ϕ[0.5 1.5] (compounded over five 200ms sub-states, this scales to [2.5 7.5] per 1 second).
CONCLUSION
CONCLUSION

In my dissertation, I showed that dopamine reward prediction errors are parsimoniously explained by a belief state TD model. I also showed that the medial prefrontal cortex shapes dopamine signals under state uncertainty, providing insight regarding the function of cortex in reinforcement learning. The belief state TD model is a critical departure from the TD models classically used to explain reinforcement learning in the brain (Schultz et al, 1997). In this conclusion, I will discuss two ideas. First, given diverse new theories for the role of dopamine in learning (Langdon et al, 2018), I will clarify exactly how the belief state TD model differs from the original model, as well as what remains the same. Second, I will discuss a possible neural implementation of the belief state. Others have proposed models for implementing a belief state in biologically plausible networks (Rao, 2004; Rao, 2005; Yu and Dayan, 2005; Zemel et al, 2005; Ma et al, 2006; Beck et al, 2008; Deneve, 2008). Coding for a belief state and efficiently conveying this information into downstream reinforcement learning circuits are interesting problems that naturally emerge from the computational framework of the belief state TD model.

First, I will clarify how the belief state TD model extends previous proposals, as well as how it maintains certain aspects of the original CSC TD model. Reinforcement learning was originally conceived as a completely ‘model-free’ system. This meant that the algorithm would try to predict all discounted future reward without an explicit model of the environment. The CSC TD model is an example of a model-free system, because it computes value over timesteps that do not derive from a model of the task, such as ISI and ITI states with associated transition probabilities and observation likelihoods. This attribute of the CSC TD model is why it produced a spurious ‘reward omission’ dip after an early reward was delivered: the model had no way of
knowing that reward signified a transition into the inter-trial interval during which no reward is expected. Any knowledge of the transition structure between states, and of the corresponding observations and probabilities of triggering these state transitions, constitute ‘model-based’ information about the structure of the task. Computing a belief state relies on model-based knowledge of the task. Therefore, the feature representation of the belief state TD model is model-based. However, the TD learning rules, including computation of the TD error and update of states’ weights, remain model free. The discrepancy between actual and predicted value \((r + \gamma \hat{V}(t + 1) - \hat{V}(t))\) is still used, identical to the CSC TD model, to signal errors and update weights. No knowledge of the task model is needed to implement this learning rule. Therefore, the belief state TD model marks a radical change in that the feature representation is model-based. However, the update rules that drive learning of state values remain model-free.

It is important to make this distinction because others have postulated that dopamine signals may signify prediction errors along axes other than value (Sharpe et al., 2017; Takahashi et al., 2017; Langdon et al., 2018). Such ‘state prediction errors’ imply that computing the error signal, thus implementing the learning rule, relies on a model of the environment. This type of prediction would ensue, for example, if the flavor of milk were switched from vanilla to chocolate (Takahashi et al., 2017). This differs from the belief state TD model, which maintains the model-free computation of the temporal difference error and update rules, and places model-based belief state computations upstream of the value function. While it is possible that model-based prediction errors drive learning about state space, it is unclear how these errors could be separated along multiple axes by downstream circuitry. Clearly, it would not benefit the organism to update the same weights for both an error in value prediction and state prediction.
As pointed out by Langdon et al, this could lead to organisms preferring endless novelty and surprises in states, for instance, over consistent quality (Langdon et al, 2018). While recent work has shown that dopamine signals broadcast different types of information to subregions of the striatum (Parker et al, 2016; Menegas et al, 2017), it remains unknown whether dopamine signals containing state prediction errors can be separated or gated downstream to drive learning about a world model. This is an important experimental hurdle that must be addressed before theories further integrate model-based prediction errors. In summary, my dissertation is fully consistent with a belief state TD model, but does not weigh in on the ongoing debate regarding model-based error computation in the dopamine system.

I will next discuss how a belief state could be computed in a neural circuit, and how these computations could be conveyed to downstream reinforcement learning circuits. Many theories exist for how neurons could represent or implement Bayesian probabilistic computations (Rao, 2004; Rao, 2005; Yu and Dayan, 2005; Zemel et al, 2005; Ma et al, 2006; Beck et al, 2008; Deneve, 2008). The hypothesis I will describe here draws a close analogy to the model used in my dissertation because it is a POMDP, yet is closer to an implementation level of analysis that could plausibly play out in a cortical circuit because it harnesses recurrent neural dynamics (Rao, 2010). In this POMDP, the belief state is computed identically to how I computed it in my dissertation:

\[
b_i(t + 1) \propto p(o(t)|i) \sum_j p(i|j) b_j(t - 1)
\]  

(1)
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 \( o(t) \) under hypothetical sub-state \( i \), and \( p(i|j) \) is the probability of transitioning from sub-state \( j \) to sub-state \( i \). It is possible to compute the belief state recursively from the belief state computed at \( (t - 1) \) because of the Markov assumption: the current state only depends on the previous state, and the current observation only depends on the current state. In other words, the previous observations can be summarized by the belief state at the previous timepoint and do not need to be explicitly incorporated into Equation 1. Based on the recursive form of this equation, Rao suggested that a recurrent neural circuit combining feedback from a previous timepoint \( b(t - 1) \) and current information about observations given state occupancy \( p(o(t)|i) \) could compute a belief state (Rao, 2010). Units within the recurrently connected layer of a simple leaky integrator network, such as the one pictured in Figure 3.1a, are modeled as having output firing rates that change over time as follows (Rao, 2010; Dayan and Abbott, 2000):

\[
\frac{dv}{dt} = -v + W \cdot u + M \cdot v
\]  

(2)

where \( W \) and \( M \) are the weight matrices of the feedforward and recurrent layers, respectively, and \( u \) and \( v \) are the firing rates of the feedforward and recurrent layers, respectively (Figure 3.1a). Written in discrete form (using firing rates at time \( t - 1 \) to compute firing rates at time \( t \)), it becomes clearer how this equation could implement the belief state computation in Equation 1 (Rao, 2010):

\[
v_t(i) = W \cdot u + \sum_{j} M(i,j)v_{t-1}(j)
\]  

(3)
where $i$ and $j$ index individual neurons in the recurrent layer. The prior probabilities $p(i|j)$, stored in the transition matrix of my computational model presented in Chapters 1 and 2, would be encoded as the weights of the recurrent matrix $M$, and the previous timestep’s belief state $b_{t-1}$ would be encoded as $v_{t-1}$. The likelihood function $p(o(t)|i)$ would be represented through the feedforward input $W \cdot u$. Although Equation 3 shows the likelihood and prior being added, rather than multiplied as in Equation 1, this could be solved by taking the logarithm of each side of the equation. In this way, a simple circuit that receives feedforward information as well as recurrent structure maintaining feedback from previous timepoints, could implement a belief state. I illustrated how my task could map onto this proposed implementation in Figure 3.1b. Individual units of the feedforward layer would convey the likelihood of each observation, given a particular state occupancy; individuals units of the output recurrent layer would each correspond to a particular state, and would each fire in proportion to the probability (or log probability) allotted to that state. Together, all units of the output layer would read out the belief state as

Figure 3.1 | Recurrent network implementation of a belief state. (a) Simple network with feedforward connections and recurrently-connected layer. $U$ and $V$ denote firing rates of feedforward and recurrently-connected layers, respectively. $W$ and $M$ denote weight matrices for feedforward and recurrent connections, respectively. Adapted from Dayan and Abbott, 2000. (b) Network schematic from (a) adapted to my task observations and states. (c) Predicted firing rates for output layer neurons in my task, if these neurons each signal one value within the belief state vector.
shown in Figure 3.1c. These output units would provide inputs to the striatum, where they could shape value predictions. The existence of these types of neural signatures—from those that represent likelihoods in the feedforward layer, to those representing components of the belief state in the recurrent layer—would be experimental predictions for future experiments.

Increasingly, an important question in reinforcement learning is how an agent knows in the first place which features, or ‘states’, to learn over. In a TD model without a belief state, the number of states being represented in the cortex and conveyed into the striatum is not constrained. Even in a simple TD model such as the CSC TD model, time could be tracked from the onset of any observable stimuli, meaning that the number of temporal ‘states’ the TD model could erroneously assign weight is enormous. Knowing which active state to assign credit for a received reward is a non-trivial problem. The belief state TD model (and the implementation I discussed above) provides an interesting lens to examine this question. In the belief state TD model, there are multiple ways in which the task dimensions could be compressed in order to solve this problem.

The first possibility is that units projecting from cortex to striatum (colored nodes in Figure 3.1b) are maximally active when they represent higher belief in a particular state—and critically, only in states that are relevant to the task. In this hypothesis, the cortex knows exactly which dimensions of computing belief are relevant to the task (Figure 3.2a). One way that the cortex may be able to accomplish this is by using spike-timing dependent synaptic mechanisms to hone a state representation based on temporal coincidences during a task (although this idea remains to be tested empirically). This was recently postulated as a mechanism for learning
about state space during vocal learning in the songbird (Mackevicius and Fee, 2018). In this proposal, Mackevicius and Fee highlight the importance of neural song ‘replay’ in premotor regions, and comparison of this replay with auditory cortical memories of tutor song. This provides a rich opportunity for assessing temporal coincidences between the two brain regions, through spike-timing dependent mechanisms and Hebbian plasticity. If the brain actively attempts to anticipate state transitions (e.g. it tries to predict reward timing, after an odor was delivered), temporal coincidences that commonly occur during the task may potentiate a particular way of representing likelihood distributions and recurrent connectivity weights in the cortex. Over learning, this would distill the belief state representation into just those needed for a particular task. Extraneous information that doesn’t show trial-by-trial temporal correlations would be automatically discarded. A second possibility is that the cortex directly uses the dopamine signal to compute a

**Figure 3.2 | Compressing belief state dimensions.** (a) Cortex computes only the relevant belief state inputs (colored nodes) to begin with. (b) Cortex computes a many-dimensional belief state with some nodes irrelevant to the task (unfilled circles). Dopamine trains a set of weights between the cortical output layer and a lower-dimensional hidden layer. Upon properly training these weights, only the belief state inputs relevant to computing value on a particular task propagate to the striatum. (c) Only the relevant belief state inputs acquire potentiated corticostriatal synapses. Dopamine could have a role in this selectivity by modulating the STDP rule.
lower-dimensional belief state that helps the animal maximize discounted future reward (Figure 3.2b). Rao proposed feeding the belief state outputs into a hidden layer containing fewer units, prior to computing value with the outputs of this small hidden layer (Rao, 2010). The weights for inputs into this ‘hidden’ layer would be tuned to the relevant belief state representation by being trained on dopamine-like TD error signals. Thus, dopamine would be used to compress the dimensions of the belief state representation, in addition to its other hypothesized role in adjusting corticostriatal weights that directly scale the value prediction. Finally, it is possible that cortex simultaneously computes many belief states (some of these irrelevant for the current task), and feeds all of these into the striatum. Then, corticostriatal plasticity is selective in such a way that only cortical inputs carrying belief states relevant to value prediction on the task achieve synaptic potentiation with their striatal targets (Figure 3.2c). One mechanism by which this could occur is dopamine-dependent modulation of the spike-timing dependent plasticity (STDP) rule (Brzosko et al, 2015). Dopamine widens the time window and temporal contingency through which STDP may occur (with dopamine present, potentiation may occur pre-post and post-pre). This provides a generous window of time during which a cortical ‘belief state’ input with imperfect temporal tuning could form a potentiated connection with a downstream striatal neuron, but only if rewards are being received (and thus dopamine is being released). The timing of dopamine release itself also modulates the strength of STDP-evoked dendritic spine enlargement in the striatum (Yagishita, 2014). In this study, dopamine release within a narrow temporal window after a pre-post pairing enhanced excitatory synaptic plasticity. In these ways, dopamine itself may play a role in strengthening only the belief state inputs into the striatum that are temporally coincident with rewards, leading to greater weights for only the subset of belief state inputs that allow the animal to maximize future rewards.
The belief state TD model extends the theoretical framework for reinforcement learning in the brain. It connects the cortex’s ability to represent diverse probabilistic models of the environment with the goal of computing accurate future values. Further experiments should probe the neural implementation of a belief state, and identify ways in which the brain efficiently knows which belief state to use in order to maximize expected future reward.
REFERENCES


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).


