HHS Public Access

Author manuscript

*Neuron*. Author manuscript; available in PMC 2019 May 02.

Published in final edited form as:

*Neuron*. 2018 May 02; 98(3): 616–629.e6. doi:10.1016/j.neuron.2018.03.036.

# Medial prefrontal cortex shapes dopamine reward prediction errors under state uncertainty

**Clara Kwon Starkweather**[1], **Samuel J. Gershman**[2,*], and **Naoshige Uchida**[1,*,**]

[1]Center for Brain Science, Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA

[2]Center for Brain Science, Department of Psychology, Harvard University, 52 Oxford Street, Cambridge, MA 02138, USA

## Abstract

Animals make predictions based on currently available information. In natural settings, sensory cues may not reveal complete information, requiring the animal to infer the 'hidden state' of the environment. The brain structures important in hidden state inference remain unknown. A previous study showed that midbrain dopamine neurons exhibit distinct response patterns depending on whether reward is delivered in 100% (Task 1) or 90% of trials (Task 2) in a classical conditioning task. Here we found that inactivation of the medial prefrontal cortex (mPFC) affected dopaminergic signaling in Task 2, in which the hidden state must be inferred ('will reward come, or not?'), but not in Task 1, where 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.

## 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

*Correspondence: N.U. uchida@mcb.harvard.edu, S.J.G. gershman@fas.harvard.edu.
**Lead Contact: Naoshige Uchida



SUPPLEMENTAL INFORMATION

Supplemental Information includes eight figures and can be found with this article online at XXX.

**AUTHOR CONTRIBUTIONS**

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

**DECLARATION OF INTERESTS**

The authors declare no competing interests.

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). Although empirical evidence has begun to support this idea as it applies to the midbrain dopamine system (Rao, 2010; Lak et al, 2017; Starkweather et al, 2017), the neural mechanisms underlying hidden state inference remain largely unknown.

The activity of dopamine neurons is sensitive to reward expectation. Dopamine neurons report a reward prediction error (RPE) signal 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 et al, 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). The magnitude of dopamine reward responses is 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 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 (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 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 1D). In a stark contrast, in Task 2, dopamine reward responses increased as time elapsed (Figure 1H). Computational modeling in which reward expectation is computed over belief states explained these divergent patterns. This model assumes transitions between two 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 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 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 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 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 expectation 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 et al, 2013; Jo and Mizumori, 2015). Moreover, these studies did not separate the contribution of interval timing 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. 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 2A and 2B) in behaving mice (Figures 2C and S1A). After subcutaneously injecting the mice with the KORD agonist salvinorin B (SalB), we observed neurons that decreased the amplitude of both baseline and task-related activity (Figures 2D, 2E, S1B, and S1C). More than 60% of recorded neurons suppressed their averaged firing rates to below half of their pre-injection rates (Figures 2F–2H). 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 2F). We analyzed any remaining task-related activity that these neurons displayed, and confirmed that task-related activity of these neurons displayed a level of suppression similar to that of the averaged recorded activity (Figure S1C). 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 (Figure S1D). Following SalB injection, neural activity no longer spanned the entire interval, instead only showing activation immediately following odor onset (Figures S1D and S1E). Taken together, KORD reliably suppressed both baseline and task-related activity, and abolished sustained activity in the mPFC during a classical conditioning task.

## Behavior and electrophysiology

We next inactivated mPFC while recording from dopamine neurons in the ventral tegmental area (VTA) (Figure 3A, Figure S2A). We injected KORD unilaterally into the mPFC of 8 mice (Figures S2B–S2F). 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 (see *Materials and Methods*, Figures 3B–3E). On each recording day, we alternated between subcutaneously injecting animals with saline (control) or SalB (mPFC inactivation) at the beginning of the session. We trained 4 mice each on Tasks 1 and 2 (Figure 3F). 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. 100% of Odor A–C trials were rewarded in Task 1, whereas 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 3G, # licks 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 S3). Animals licked more (# anticipatory licks for Odor A in Task 1 ≠ Task 2; $F_{1,77} = 6.7$, $p = 1.2 \times 10^{-2}$, ANOVA; Figure S3A), 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 ANOVA), in Task 1 compared to Task 2 (Figure 3G, Figures S3G and S3H). In general, animals ramp up their lick rates soon for higher reward probabilities, even in the absence of variability in the timing of reward (Fiorillo et al, 2003; Tian and Uchida, 2015). SalB did not affect any measures used to quantify the pattern of licking across time (Figure S3B–S3I).

We asked whether the types of neurons recruited to the task differed between Saline and SalB conditions. We applied k-means clustering to all of our recorded VTA neurons (those within 500um of an optogenetically-identified dopamine neuron; n = 761 neurons) and sorted neurons into three clusters that showed phasic activity to cue and reward, sustained positive activity, and sustained negative activity (Figure S4A) (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 recorded neurons, between Saline and SalB conditions in each task (Figure S4B).

### 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 reward responses following Odor A showed opposite trends of temporal modulation between Tasks 1 and 2, replicating our earlier results (Starkweather et al, 2017). In Task 1, post-reward responses decreased as a function of time (Figure 4A, 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 4B, colored lines; $F_{8,320} = 3.7$, $p = 3.8 \times 10^{-4}$, 2-way ANOVA; factors: ISI, neuron). Scalar timing uncertainty could not account for the positive temporal modulation of post-reward responses for Odor A. The post-reward response to Odor C (with an ISI of 2.8s, see Figure 3F), was significantly smaller than the post-reward response to the latest possible Odor A reward, which also had an ISI of 2.8s (Figure 4B; $Odor\ A_{ISI=2.8s} \neq Odor\ C\ response$, $F_{1,41} = 22.4$, $p = 2.7 \times 10^{-5}$, 2-way ANOVA; factors: ISI, neuron). In addition, pre-reward firing rates decreased over time (Figures 4A and 4B, 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 decreased as a function of time, similar to the control data (Figure 4C, 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 4D, colored lines; $F_{8,368} = 0.69$, $p = 0.70$, 2-way ANOVA; factors: ISI, neuron) although the responses were still smaller (~55%) compared to their responses to unexpected reward (Figure S5B). Interestingly, pre-reward RPEs in both Tasks 1 and 2 decreased over time, similar to the control condition (Figures 4C and 4D, 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 due to KORD inefficacy in Task 1, we simultaneously recorded mPFC neurons in one Task 1 animal (Figures S6A and S6B). We confirmed that KORD inhibited the majority of mPFC neurons in this Task 1 animal (Figures S6C–S6E). 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 (Figures S6F–S6H). In summary, mPFC inactivation selectively abolished temporal modulation of post-reward responses in Task 2, in which reward was delivered probabilistically (Figure 4E).

To 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 ($t_r$) × Task 1, $t_r$ × Task 2, $t_r$ × Drug × Task 1, $t_r$ × Drug × Task 2, and free water response. Free water response was included because dopamine neurons show great variability in the magnitude of reward responses (Eshel et al., 2015). 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 ($t_r$ × Task 1: $F_{8,1347} = 2.2$, $p = 2.7 \times 10^{-2}$; $t_r$ × Task 2: $F_{8,1347} = 3.2$, $p = 1.1 \times 10^{-3}$; $t_r$ × drug × Task 2: $F_{9,1347} = 2.6$, $p = 6.2 \times 10^{-3}$; $t_r$ × drug × Task 1: $F_{9,1347} = 0.79$, $p = 0.62$). 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 post-reward spikes on every Odor A trial (Figures 5A–5D). 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 5F). In contrast, 24 out of 47 (51%) displayed negative slopes in the inactivation condition (Figure 5F). In Task 1, the proportion of significant slopes showing negative versus positive modulation was not significantly different between drug and control conditions (Figure 5E; Saline − 16 negative: 0 positive; SalB − 16 negative, 2 positive, $p = 0.49$, Fisher exact test), whereas in Task 2, the proportions significant slopes showing negative versus positive modulation was significantly altered (Figure 5F; Saline − 1 negative: 9 positive; SalB − 9 negative: 7 positive, $p = 4.1 \times 10^{-2}$, Fisher exact test). Therefore, neurons were more likely to show significant negative temporal modulation in Task 2, in the SalB condition.

Accordingly, the distribution of slopes in the inactivation condition shifted significantly towards smaller values in Task 2 (Figure 4E, Figure 5F; $median_{Saline} \neq median_{SalB}$, $z = 2.1$, $p = 3.8 \times 10^{-2}$, Wilcoxon rank-sum test). The distribution of slopes was not different in the SalB condition in Task 1 (Figure 5E; $median_{Saline} \neq median_{SalB}$, $z = 0.14$, $p = 0.89$, Wilcoxon rank-sum test). To assess the effect of task and drug on all neurons' slopes, we performed an ANOVA that included the following factors: Task, Task 1 × Drug, and Task 2 × Drug. To eliminate deviation from the normal distribution in groups, we normalized post-reward responses to free water responses prior to computing slopes. Task, and Task 2 × Drug, both explained a significant level of variance in the slopes (Task: $F_{1,163} = 2.2$, $p = 1.7 \times 10^{-5}$; Task 2 × Drug: $F_{1,163} = 4.7$, $p = 3.2 \times 10^{-2}$), while Task 1 × Drug did not explain a significant level of variance in the slopes (Task 1 × Drug: $F_{1,163} = 6.5 \times 10^{-2}$, $p = 0.94$). 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 5F; $\sigma^2_{Saline} \neq \sigma^2_{SalB}$, $F_{40,46} = 0.51$, $p = 0.03$, $F$-test for equality of two variances), with increased variance in the inactivation versus the control condition being reflected in individual animals (Figure 5F, black dots). The variance of the slopes was not significantly greater in the inactivation condition, in Task 1 (Figures 5E; $\sigma^2_{Saline} \neq \sigma^2_{SalB}$, $F_{41,41} = 0.59$, $p = 0.10$, $F$-test for equality of two variances). Therefore, the abolished temporal modulation shown in Figure 4D for Task 2 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.

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. 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. Post-reward RPEs are greater for longer ISIs due to scalar timing noise (Fiorillo et al, 2008; Jo and Mizumori, 2015; Kobayashi and Schultz, 2008). Consistent with this, we found that post-reward RPEs following Odor C were larger than those following Odor B (Figures 6A and 6B). 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 was not larger during mPFC inactivation. On reward omission trials, dopamine neurons briefly paused their tonic firing rates at the time of expected reward (Figures 6C and 6D, Figure S5D). 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, consistent with a previous study (Jo and Mizumori, 2015). In summary, mPFC inactivation impaired positive temporal modulation of post-reward responses in Task 2, but spared 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

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 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; 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 7A). 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.

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 7A, parameters in Figures S7A and S7B) 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. The model allots 100% of its belief, sequentially, into each of the ISI sub-states, as time elapses during the trial (Figure 7B, S8A, and S8B). In contrast, Task 2 is a hidden Markov decision process because the cue may lead to an omission trial in 10% of cases (Figure 7A, parameters in Figures S7A and S7C). 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 7B, S8C, and S8D). 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 7B, S8C, and S8D).

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 7C, left panel) because the model learned higher weights for later features, reflecting the higher momentary probability of receiving reward at later timepoints (Figures S8A and S8B). Post-reward firing in Task 2 increased over time (Figure 7C, right panel), reflecting the model's mounting belief that it is in an omission trial at later timepoints (Figures S8C and S8D). 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 5. This revealed that the majority of simulations produced positive temporal modulation over time (Figure 7D). Pre-reward firing decreased over time in both Tasks 1 and 2 (Figure 7C, black lines). Finally, our model produced omission responses around the time of expected reward, for Odors B and C (Figure 7E).

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 S7E), 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 7B) was fixed with 100% probability in the ISI, and remained uniformly 'stuck' at this probability, similar to the Task 1 belief state (Figure S8E). 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. In a model with an intact belief state, the ITI state accrues greater probability as time elapses, and accordingly, the sum of the probability allotted to the ISI sub-states decreases over time (Figure S8D). These probabilities do not change over time in the impaired model (Figure S8E). 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 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 7F and Figure 8). Similar to 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 7G, compare with intact model in Figure 7D), similar to our inactivation data (Figure 5F). These negatively modulated post-reward RPEs occurred in simulations with the corrupted belief state (Figures 8 and S8E). Finally, simulations run with a deficit in hidden state inference still exhibited a reward omission response around the time of expected reward (Figure 7H) because later sub-states accrued very low weights (see weights in Figures S8C and S8D), 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 examined the effect of impairing the model's timing mechanism by blurring the transition probabilities between sub-states (Takahashi et al, 2016, Figure S7D). This timing impairment could somewhat flatten temporal modulation of Task 2 post-reward responses (Figure 7I, 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 7I, left panel), pre-reward responses in both Tasks 1 and 2 were flattened (Figure 7I, black lines), most simulations still showed positive modulation of post-reward responses over time (Figure 7J), and reward omission responses were abolished (Figure 7K).

We attempted to blur the model's timing estimation less dramatically than in Figure S7D, but this increased the positive temporal modulation of post-reward responses in Task 2. This occurred because a smaller increase in scalar timing noise makes only later rewards harder to predict, thereby increasing positive temporal modulation. In contrast, the parameter 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 uniformly blur the transition matrix. However, this eliminated reward omission responses, inconsistent with our data. Our data are most consistent with mPFC shaping hidden state inference, rather than timing estimation, in the dopaminergic circuitry.

## DISCUSSION

Our results demonstrate that the sensitivity of dopamine RPEs to state uncertainty can be explained by a computational framework that incorporates a belief state. We 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), these studies did not disentangle the involvement of mPFC in time

estimation *versus* in an inferential process that evolves across time. Kim et al 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 (categorizing 'short' versus 'long') were flattened, post-mPFC inactivation. This task can be conceptualized as requiring hidden state inference. Based on the time interval, the animal infers the correct (hidden) categorization. As time elapses during the interstimulus interval, the animals' belief state increasingly favors 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. 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. How is interval timing information conveyed to dopamine neurons? 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 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 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 the belief state be conveyed from mPFC to dopamine neurons? Several routes exist. First, the mPFC sends ipsilateral projections 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. For example, the mPFC sends dense ipsilateral projections to the striatum, including the nucleus accumbens (Sesack, et al, 1989; Vertes, 2004; Gabbott et al, 2005), which then supplies major inputs to dopamine neurons in the VTA (Watabe-Uchida, 2012). 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 belief state information reaches dopamine neurons on the recorded hemisphere, accounting for why we observed a partial effect of inactivation (we ran 60% of simulations with an impaired belief state to fit our data).

Previous studies have implicated the orbitofrontal cortex (OFC) in representing state space (Takahashi et al, 2011; Bradfield et al, 2011). Our mPFC inactivation produced a different, but related, impairment. Rather than ablating the state representation, 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. A basic representation of observable state space remains intact upon mPFC inactivation. If the state representation were abolished, the brain could no longer distinguish the ISI from the ITI, even 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). The RPEs produced by this TD model would simply reflect the temporal distribution of rewards, reproducing a Gaussian distribution in both Tasks 1 and 2 (see Starkweather et al, 2017). For this reason, our results are incompatible with an ablated state representation. 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. The mPFC then computes a probability distribution over these possible hidden states furnished by the OFC.

Another remaining question is how a state representation is formed in the first place. Computational hypotheses addressing this question exist (Gershman et al, 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. One idea is that the brain identifies temporally informative stimuli in the environment, and forms a state representation based on this. A region attuned to detecting these coincidences could be the hippocampus, theorized to generate a predictive map that respects the transition structure within a (usually spatial) task (Stachenfeld et al, 2017). Such a predictive map, if applied instead to transitions in time rather than space, could build a useful state representation in a classical conditioning task with temporally informative cues. Recently, several studies have suggested that the hippocampus stores such temporal relationships between stimuli (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 stimuli. Based on this map, task states for reinforcement learning could be formed 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. 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 (Ludvig et al, 2008). 'State reset' TD models 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 S8B); in Task 2, these weights decrease for later timepoints, similar to the value function in Task 2 (Figure S8D). 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, which would have to re-learn a new set of weights in order to switch their response patterns. Rather, our inactivation is parsimoniously explained by hidden state inference being selectively and transiently abolished by our experimental manipulation.

Could our data be explained by mPFC encoding the risk of a trial going unrewarded? 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 Odor B and C trials were no longer risky, expectation should be higher in the mPFC inactivation condition, resulting in more suppressed reward responses. 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 S5C). 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 S5C). In other words, RPEs were bigger in riskier conditions than in non-risky conditions, regardless of whether the mPFC was inactivated or not. Therefore, broadly eliminating risk-related information cannot explain our data.

How might the mPFC represent the neural correlates of a belief state? Our mPFC inactivation spanned subregions of the mPFC, including both prelimbic (PL) and infralimbic (IL) cortices (Figure S2). 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, the belief state must favor a rewarded state. Conversely, to extinguish a former reward-predicting cue, the 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 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 colored bars for Task 2 in Figure 7B), while IL neurons show ramping that increases as time elapses (similar to black bars for Task 2 in Figure 7B). 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 belief in the rewarded state, RPEs in the 90%-rewarded task should become larger and tend towards more positive temporal modulation, due to the model favoring reward omission. If IL were lesioned, RPEs should become smaller and favor negative temporal modulation. These hypothetical differences between PL and IL could also produce corresponding behavioral phenotypes. Behavioral paradigms aiming to capture modulation by a belief state should incentivize the animal to behave differently once

the belief state shifts. Li and Dudman trained mice on an operant task in which rewards were delivered in a Gaussian distribution of ISIs (Li and Dudman, 2013). In probe trials, rewards were not given. Upon lesioning PL—potentially involved in signaling the rewarded state—mice might wait in the reward port for a shorter duration prior to giving up and re-initiating a trial, whereas the opposite effect would be expected upon lesioning IL. In future work, it would be important to characterize the activities of these various subregions of the mPFC.

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.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Surgery and viral injections
  - Behavioral paradigm
  - SalB injection
  - Electrophysiology
  - Immunohistochemistory
  - Computational modeling
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Data analysis
  - Statistics
- DATA AND SOFTWARE AVAILABILITY
  - Code availability
  - Data availability
- KEY RESOURCES TABLE

## STAR ★METHODS

### CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by Clara Starkweather (skaralc@gmail.com) & Naoshige Uchida (uchida@mcb.harvard.edu).

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

We used 12 adult male mice ranging in age from 6 to 24 months of age, heterozygous for the transgene that expresses Cre recombinase under the control of the DAT promoter (B6.SJL-Slc6a3$^{tm1.1(cre)Bkmm}$/J, The Jackson Laboratory, Backman et al, 2007), backcrossed for at least five generations with C57/BL6 mice. Animals ranged in weight from 20–25g. 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. Animals were singly 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, between 10AM and 7PM. 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.

## METHOD DETAILS

**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).

**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. 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. 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 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 = 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 (Bransonic). 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. We analyzed data recorded in the 40 minutes following recording onset, as this corresponded to the length of time we confirmed neural activity to be suppressed by KORD (Figure 2F).

**Electrophysiology—**We based recording techniques on previous studies (Cohen et al, 2012; Tian and Uchida, 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/mm$^2$ 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 and Uchida, 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 (<3ms) jitter in spike latency. While others have used a latency criteria of 5ms or less ('short latency') (Cohen et al, 2012; Tian and Uchida, 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.

**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 (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 S2A), and to verify that the KORD expression was in the mPFC (Figures S2B–S2F).

## 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 S7 and S8), 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 7A). 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:

$$b_i(t+1) \propto p(o(t)\,|\,i) \sum_j p(i\,|\,j) b_j(t-1) \quad (1)$$

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 $o(t) \in \{cue, reward, null\}$ 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):

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

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 $b(t)$:

$$\hat{V}(t) = \sum_i w_i b_i(t) \quad (3)$$

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

$$\Delta w_i = \alpha b_i(t) \delta(t) \quad (3)$$

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

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

In the belief state TD model, it is assumed that the animal has learned a state transition distribution, encoded by matrix T (Figure S7A). 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→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→7. We captured the exponential distribution of dwell-times in the ITI state by setting T (30, 30) to 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 S7B and S7C):

Task 1:

$$\mathrm{T}(30, 30) = 1 - \mathrm{ITI\_hazard}$$

$$\mathrm{T}(30, 1) = \mathrm{ITI\_hazard}$$

Task 2:

$$\mathrm{T}(30, 30) = 1 - \mathrm{ITI\_hazard} * 0.9$$

$$\mathrm{T}(30, 1) = \mathrm{ITI\_hazard} * 0.9$$

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 S7A). 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:

$$\mathrm{O(i, j, 1)} = \text{likelihood of observation of 'null', given i} \rightarrow \text{j transition}$$

$$\mathrm{O(i, j, 2)} = \text{likelihood of observation of 'cue', given i} \rightarrow \text{j transition}$$

$$O(i, j, 3) = \text{likelihood of observation of 'reward', given i} \rightarrow \text{j transition}$$

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 S7B and S7C):

Task 1:

$$O(30, 30, 1) = 1(\text{null observation})$$

Task 2:

$$O(30, 30, 1) = 1 - \text{ITI\_hazard} * 0.1(\text{null observation})$$

$$O(30, 30, 2) = \text{ITI\_hazard} * 0.1(\text{cue in a small percentage of cases})$$

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.

**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:

$$\tilde{p}(t) = \frac{1}{\phi t\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\tau) e^{\frac{-(\tau - t)^2}{2\phi^2 t^2}} d\tau \quad (5)$$

We used a Weber fraction $\phi = 0.05$ (compounded over five 200ms sub-states, this scales to 0.25 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 (Figures S3J–S3L). 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 500 sessions, consisting of 50 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 4A), 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. 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:

$$O = (30, 30, 2) = \text{ITI\_hazard} * 0.1(\text{cue in a small percentage of cases})$$

$$O(30, 30, 1) = 1 - \text{ITI\_hazard} * 0.1(\text{null observation})$$

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 8 and S8E)

**Impairing timing:** We impaired timing by increasing the Weber fraction used to blur the transition matrix. The simulations shown in Figures 7I–K used a range of Weber fractions $\phi$[0.5 1.5] (compounded over five 200ms sub-states, this scales to [2.5 7.5] per 1 second).

## QUANTIFICATION AND STATISTICAL ANALYSIS

**Data analysis—**The values reported in the text and error bars are the mean ± s.e.m. unless otherwise noted. All data was analyzed and modeled with MATLAB 2015a (Mathworks). Statistical tests were performed in MATLAB. Nonparametric tests were used where appropriate (if a chi-square goodness of fit test indicated deviance from a normal distribution) and tests were 2-tailed. Alpha was pre-set to 0.05. The first author was not blinded to the experimental conditions. Sample size was not predetermined, but the number of mice per group matches similar studies using optogenetic identification of dopamine neurons in mice (Cohen et al, 2102; Tian and Uchida, 2015; Eshel et al, 2015).

We focused our analysis on light-identified dopamine neurons. We analyzed data recorded in the 40 minutes following recording onset, as this corresponded to the length of time we confirmed neural activity to be suppressed by KORD (Figure 2F). 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 and Uchida, 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. For ANOVAs in which Task and Drug were factors, 'Task 1', 'Task 2', and 'drug' had values of 1 or 0 depending on the task, and whether or not drug was present.

## DATA AND SOFTWARE AVAILABILITY

**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

**Data Availability—**The data that support the findings of this study are available online at https://crcns.org/ (to be added once request for deposit is approved)

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We thank M. Watabe-Uchida, M. Krashes, C. Li, R. Born, J. Assad, J. Paton, and C. Harvey for discussions. We thank V. Murthy and C. Dulac for sharing resources. This work was supported by National Science Foundation grant CRCNS 1207833 (S.J.G.), National Institutes of Health grants R01MH095953 (N.U.), R01MH101207 (N.U.), F30MH112242-01A1 (C.K.S.), T32MH020017 (C.K.S.), T32GM007753 (C.K.S.), Harvard Brain Science Initiative Seed Grant (N.U.), the Simons Collaboration on Global Brain (N.U.), and Harvard Mind Brain and Behavior faculty grant (S.J.G. and N.U.).

## References

Atasoy D, Aponte Y, Su HH, Sternson SM. A FLEX switch targets channelrhodopsin-2 to multiple cell types for imaging and long-range circuit mapping. J. Neurosci. 2008; 28:7025–7030. [PubMed: 18614669]

Backman C, Malik N, Zhang Y, Shan L, Grinberg A, Hoffer B, Westphal H. Characterization of a mouse strain expressing Cre recombinase from the 30 untranslated region of the dopamine transporter locus. Genesis. 2007; 45:418–426. [PubMed: 17549727]

Balsam PD, Gallistel CR. Temporal maps and informativeness in associative learning. Trends Neurosci. 2009; 32:73–78. [PubMed: 19136158]

Bayer HM, Glimcher PW. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron. 2005; 47:129–141. [PubMed: 15996553]

Bradfield L, Dezfouli A, van Holstein M, Chieng B, Balleine B. Medial Orbitofrontal Cortex Mediates Outcome Retrieval in Partially Observable Task Situations. Neuron. 2011; 88:1268–1280.

Cohen JY, Haesler S, Vong L, Lowell BB, Uchida N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature. 2012; 482:85–88. [PubMed: 22258508]

Carr DB, Sesack SR. Projections from the rat prefrontal cortex to the ventral tegmental area: target specificity in the synaptic associations with mesoaccumbens and mesocortical neurons. J. Neurosci. 2000; 20:3864–3873. [PubMed: 10804226]

Courville AC, Daw ND, Touretzky DS. Bayesian theories of conditioning in a changing world. Trends Cogn. Sci. 2006; 10:294–300. [PubMed: 16793323]

Daw ND, Courville AC, Tourtezky DS. Representation and timing in theories of the dopamine system. Neural Comput. 2006; 18:1637–77. [PubMed: 16764517]

Deuker L, Bellmund J, Schroder T, Doeller C. An event map of memory space in the hippocampus. eLife. 2016; 5:e16534. [PubMed: 27710766]

Eshel N, Bukwich M, Rao V, Hemmelder V, Tian J, Uchida N. Arithmetic and local circuitry underlying dopamine prediction errors. Nature. 2015; 525:243–246. [PubMed: 26322583]

Fiorillo CD, Tobler PN, Schultz W. Discrete coding of reward probability and uncertainty by dopamine neurons. Science. 2003; 299:1898–1902. [PubMed: 12649484]

Fiorillo CD, Newsome WT, Schultz W. The temporal precision of reward prediction in dopamine neurons. Nat. Neurosci. 2008; 11:966–973. [PubMed: 18660807]

Fiser J, Berkes P, Orbán G, Lengyel M. Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn. Sci. 2010; 14:119–130. [PubMed: 20153683]

Gabbott PLA, Warner TA, Jays PRL, Salway P, Busby SJ. Prefrontal cortex in the rat: projections to subcortical autonomic, motor, and limbic centers. J. Comp. Neurol. 2005; 492:145–177. [PubMed: 16196030]

Gershman SJ, Blei DM, Niv Y. Context, learning, and extinction. Psychol. Rev. 2010; 117:197–209. [PubMed: 20063968]

Gershman SJ, Norman KA, Niv Y. Discovering latent causes in reinforcement learning. Curr. Opin. Behav. Sci. 2015; 5:43–50.

Gourley S, Taylor J. Going and stopping: dichotomies in behavioral control by the prefrontal cortex. Nat. Neurosci. 2016; 19:656–664. [PubMed: 29162973]

Guo Z, Inagaki H, Daie K, Druckmann S, Gerfen CR, Svoboda K. Maintenance of persistent activity in a frontal thalamocortical loop. Nature. 2017; 545:181–186. [PubMed: 28467817]

Howard MW, Eichenbaum H. Time and space in the hippocampus. Brain Res. 2015; 1621:345–354. [PubMed: 25449892]

Janssen P, Shadlen MN. A representation of the hazard rate of elapsed time in macaque area LIP. Nat. Neurosci. 2005; 8:234–241. [PubMed: 15657597]

Jo YS, Mizumori SJY. Prefrontal regulation of neuronal activity in the ventral tegmental area. Cereb. Cortex. 2015; 26:4057–4068. [PubMed: 26400913]

Jo YS, Lee J, Mizumori SJY. Effects of prefrontal cortical inactivation on neural activity in the ventral tegmental area. J. Neurosci. 2013; 33:8159–8171. [PubMed: 23658156]

Kim J, Ghim J-W, Lee JH, Jung MW. Neural correlates of interval timing in rodent prefrontal cortex. J. Neurosci. 2013; 33:13834–47. [PubMed: 23966703]

Kim J, Jung AH, Byun J, Jo S, Jung MW. Inactivation of medial prefrontal cortex impairs time interval discrimination in rats. Front. Behav. Neurosci. 2009; 3:38. [PubMed: 19915730]

Kobayashi S, Schultz W. Influence of reward delays on responses of dopamine neurons. J. Neurosci. 2008; 28:7837–7846. [PubMed: 18667616]

Kvitsiani D, Ranade S, Hangya B, Taniguchi H, Huang JZ, Kepecs A. Distinct behavioural and network correlates of two interneuron types in prefrontal cortex. Nature. 2013; 498:363–366. [PubMed: 23708967]

Lak A, Nomoto K, Keramati M, Sakagami M, Kepecs A. Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. Curr. Biol. 2017; 27:821–32. [PubMed: 28285994]

Li Y, Dudman JT. Mice infer probabilistic models for timing. Proc. Natl. Acad. Sci. 2013; 110:17154–17159. [PubMed: 24082097]

Lima SQ, Hromádka T, Znamenskiy P, Zador AM. PINP: A new method of tagging neuronal populations for identification during in vivo electrophysiological recording. PLoS ONE. 2009; 4:e6099. [PubMed: 19584920]

Ludvig EA, Sutton RS, Kehoe EJ. Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. Neural Comput. 2008; 20:3034–3054. [PubMed: 18624657]

Mello G, Soares S, Paton J. A scalable population code for time in the striatum. Curr. Biol. 2015; 25:1113–1122. [PubMed: 25913405]

Mitchell AS. The mediodorsal thalamus as a higher order thalamic nucleus important for learning and decision-making. Neurosci. Biobehav. Rev. 2015; 54:76–88. [PubMed: 25757689]

Nomoto K, Schultz W, Watanabe T, Sakagami M. Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. J. Neurosci. 2010; 30:10692–10702. [PubMed: 20702700]

Oprisan SA, Aft T, Buhusi M, Buhusi C. Scalar timing in memory: A temporal map in the hippocampus. J. Theor. Biol. 2018; 438:133–142. [PubMed: 29155279]

Pasquereau B, Turner RS. Dopamine neurons encode errors in predicting movement trigger occurrence. J. Neurophysiol. 2015; 113:1110–1123. [PubMed: 25411459]

Pouget A, Beck JM, Ma WJ, Latham PE. Probabilistic brains: knowns and unknowns. Nat. Neurosci. 2013; 16:1170–1178. [PubMed: 23955561]

Rao RPN. Decision making under uncertainty: a neural model based on partially observable markov decision processes. Front. Comput. Neurosci. 2010; 4:146. [PubMed: 21152255]

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275:1593–1599. [PubMed: 9054347]

Sesack SR, Deutch AY, Roth RH, Bunney BS. Topographical organization of the efferent projections of the medial prefrontal cortex in the rat: an anterograde tract-tracing study with Phaseolus vulgaris leucoagglutinin. J. Comp. Neurol. 1989; 290:213–242. [PubMed: 2592611]

Shusterman R, Smear M, Koulakov A, Rinberg D. Precise olfactory responses tile the sniff cycle. Nat. Neurosci. 2011; 14:1039–1044. [PubMed: 21765422]

Stachenfeld K, Botvinick M, Gershman S. The hippocampus as a predictive map. Nat Neurosci. 2017; 20:1643–1653. [PubMed: 28967910]

Starkweather CK, Babayan BM, Uchida N, Gershman SJ. Dopamine reward prediction errors reflect hidden-state inference across time. Nat. Neurosci. 2017; 20:581–589. [PubMed: 28263301]

Suri R, Schultz W. Learning of sequential movements by neural network model with dopamine-like reinforcement signal. Exp. Brain Res. 1998; 121:350–354. [PubMed: 9746140]

Sutton RS. Learning to predict by the methods of temporal differences. Mach. Learn. 1988; 3:9–44.

Takahashi Y, Langdon A, Niv Y, Schoenbaum G. Temporal Specificity of Reward Prediction Errors Signaled by Putative Dopamine Neurons in Rat VTA Depends on Ventral Striatum. Neuron. 2016; 91:182–193. [PubMed: 27292535]

Takahashi Y, Roesch M, Wilson R, Toreson K, O'Donnell T, Niv Y, Schoenbaum G. Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. Nat. Neurosci. 2011; 14:1590–1597. [PubMed: 22037501]

Tian J, Uchida N. Habenula lesions reveal that multiple mechanisms underlie dopamine prediction errors and prediction error-based learning. Neuron. 2015; 87:1304–1316. [PubMed: 26365765]

Tsunoda Y, Kakei S. Reaction time changes with the hazard rate for a behaviorally relevant event when monkeys perform a delayed wrist movement task. Neurosci. Lett. 2008; 433:152–157. [PubMed: 18243554]

Uchida N, Mainen ZF. Speed and accuracy of olfactory discrimination in the rat. Nat. Neurosci. 2003; 6:1224–1229. [PubMed: 14566341]

Vardy E, et al. A New DREADD Facilitates the Multiplexed Chemogenetic Interrogation of Behavior. Neuron. 2015; 86:936–946. [PubMed: 25937170]

Vertes RP. Analysis of projections from the medial prefrontal cortex to the thalamus in the rat, with emphasis on nucleus reuniens. J. Comp. Neurol. 2001; 442:163–187.

Vertes RP. Differential projections of the infralimbic and prelimbic cortex in the rat. Synapse. 2004; 51:32–58. [PubMed: 14579424]

Watabe-Uchida M, Zhu L, Ogawa SK, Vamanrao A, Uchida N. Whole-brain mapping of direct inputs to midbrain dopamine neurons. Neuron. 2012; 74:858–873. [PubMed: 22681690]

Wilson R, Takahashi Y, Schoenbaum G, Niv Y. Orbitofrontal cortex as a cognitive map of task space. Neuron. 2014; 81:267–279. [PubMed: 24462094]

Xu M, Zhang S, Dan Y, Poo M. Representation of interval timing by temporally scalable firing patterns in rat prefrontal cortex. Proc. Natl. Acad. Sci. 2014; 111:480–485. [PubMed: 24367075]

## HIGHLIGHTS

- Dopamine reward prediction errors (RPEs) reflect hidden state inference
- Medial prefrontal cortex (mPFC) shapes RPEs in a task involving hidden states
- mPFC is not needed to compute RPEs in a similar task when states are fully observable
- Modeling suggests that mPFC computes a probability distribution over hidden states

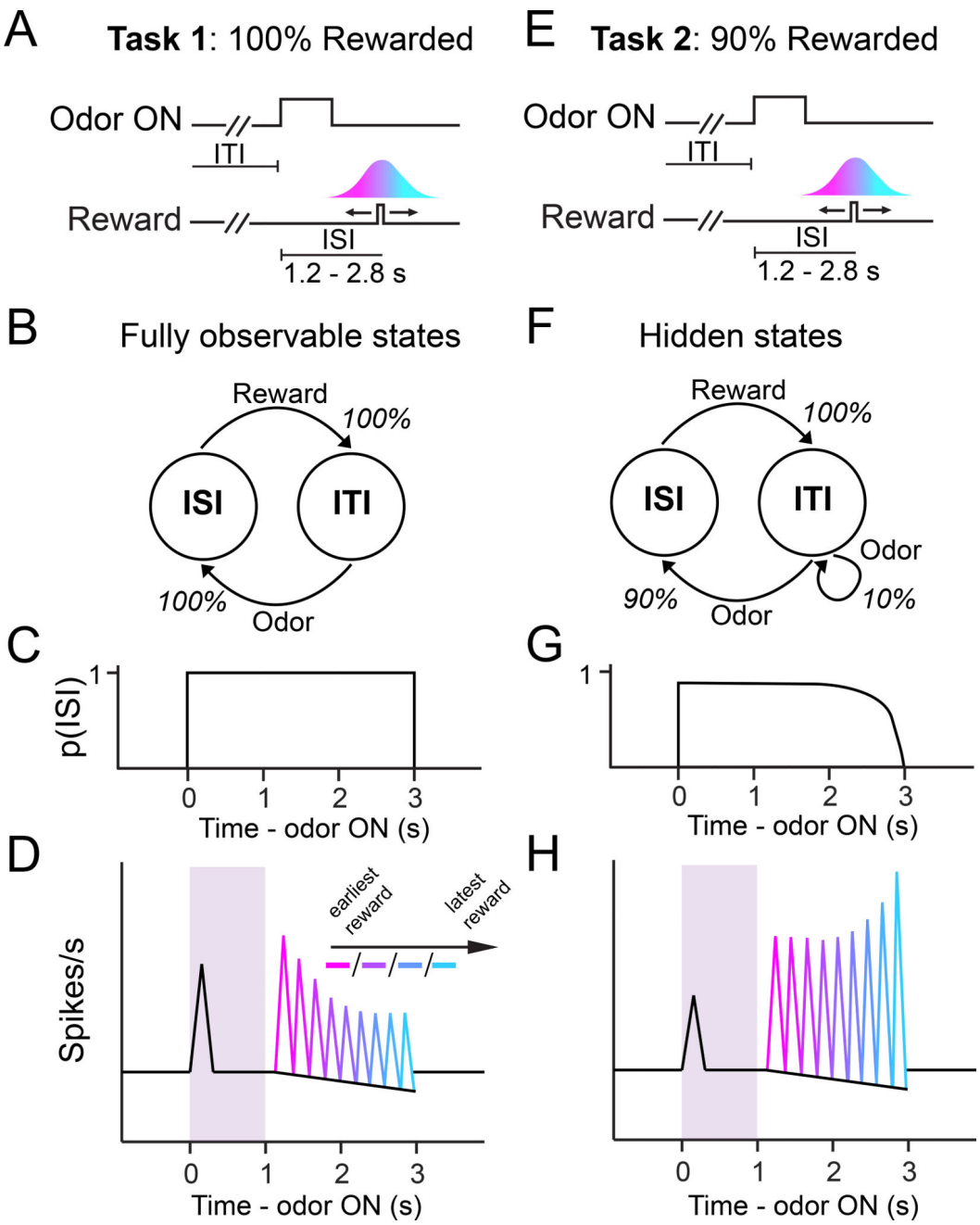


**Figure 1. Classical conditioning tasks that vary reward timing produce divergent patterns of dopamine responses, depending on whether reward is delivered deterministically**

(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' (ISI or ITI) 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 (Starkweather et al, 2017).

(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 (Starkweather et al, 2017).

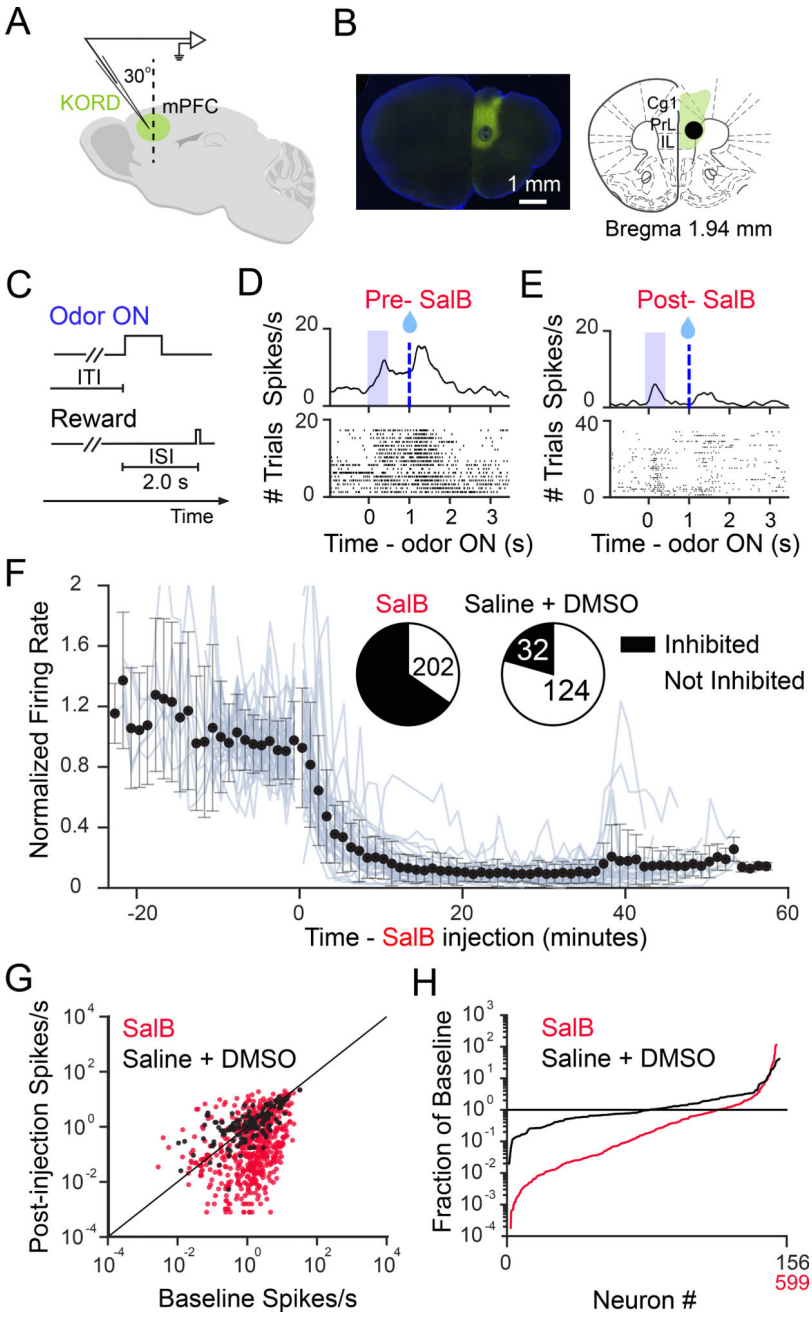


**Figure 2. KORD inactivates mPFC neurons**

(A) Tetrodes were implanted into the mPFC, in 2 mice injected with KORD in the mPFC.

(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.

(C–E) Task and firing patterns for a representative neuron, before and after SalB injection.

(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. Neurons below the unity line had lower firing rates after 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. Neurons below the horizontal line had lower firing rates after injection.

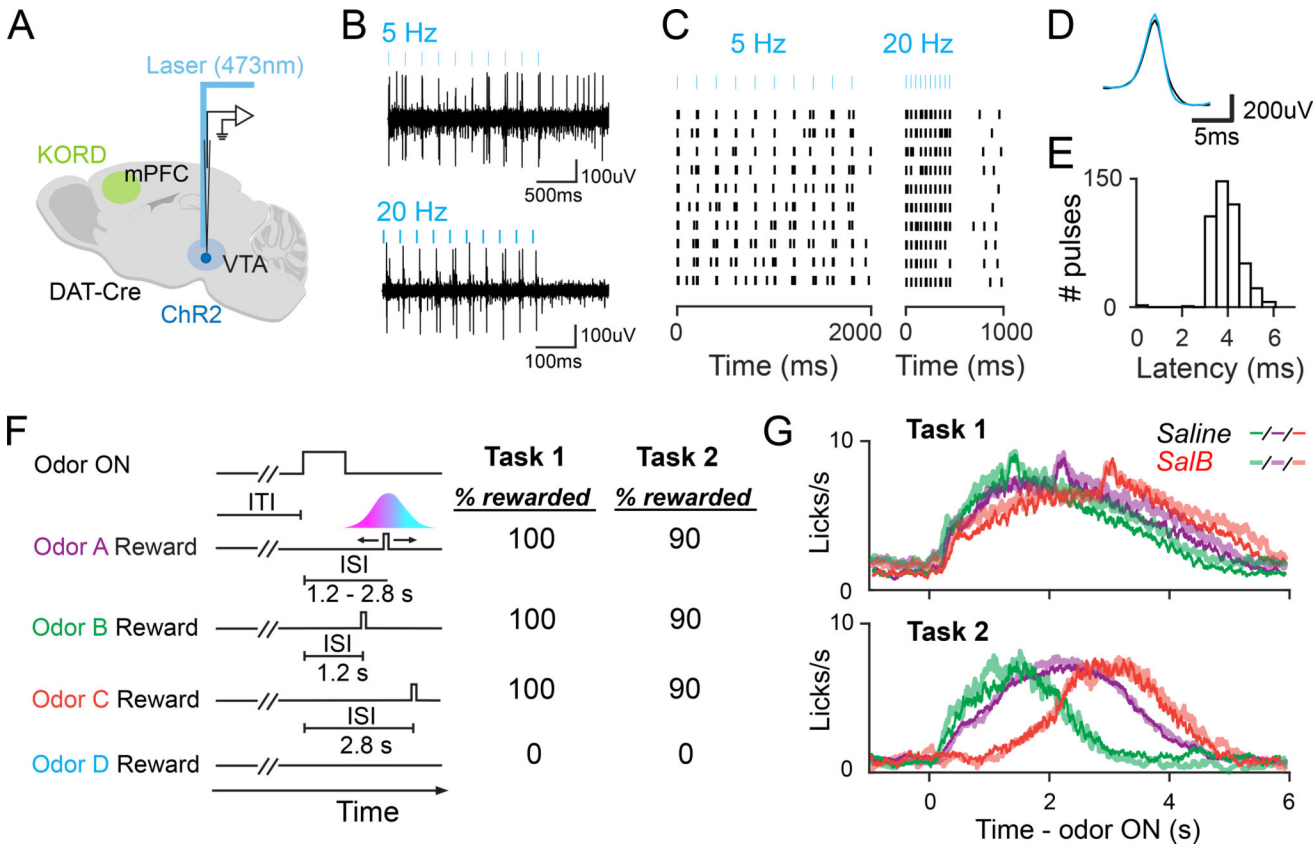


**Figure 3. Dopamine electrophysiology and behavior**

(A) Tetrodes were implanted into VTA, with KORD expression in mPFC. Neurons were included in the dataset if they responded with short latency to laser pulses.

(B–C) Raw traces and light-evoked spikes for a dopamine neuron, demonstrating laser pulses and laser-evoked spikes.

(D) Near-identical match between waveform of average laser-evoked spike (blue) and average spontaneous spike (black) for same example neuron.

(E) Histogram of latencies to spike for same example neuron.

(F) Following Odor A, reward was delivered after a variable delay time ranging from 1.2s to 2.8s. Following Odors B and C, reward was delivered after a constant delay time of 1.2s and 2.8s, respectively. Odor D was not rewarded.

(G) Averaged licking histograms for all KORD-expressing animals trained on Tasks 1 and 2, during Saline (thin, opaque lines) and SalB (thick, transparent lines) sessions. Only Task 1 Odor A trials with an ISI of 2s are shown here. All Task 2 trials included in this plot are reward omission trials.

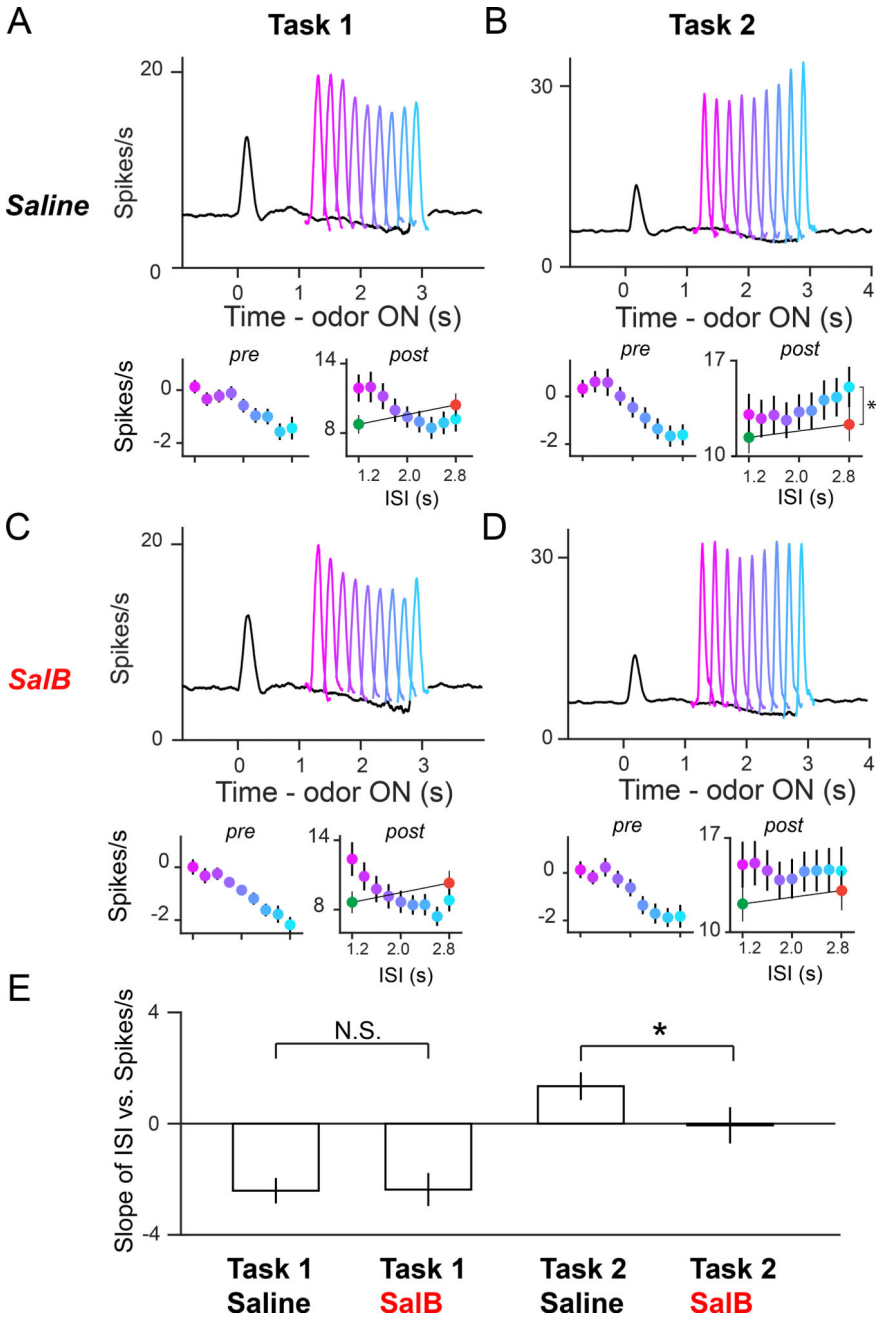


**Figure 4. mPFC inactivation impaired temporal modulation of dopamine reward responses in a non-deterministically rewarded task (Task 2) but not in a deterministically rewarded task (Task 1)**

(A) PSTH for 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.

(B) PSTH for 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 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) PSTH for 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).

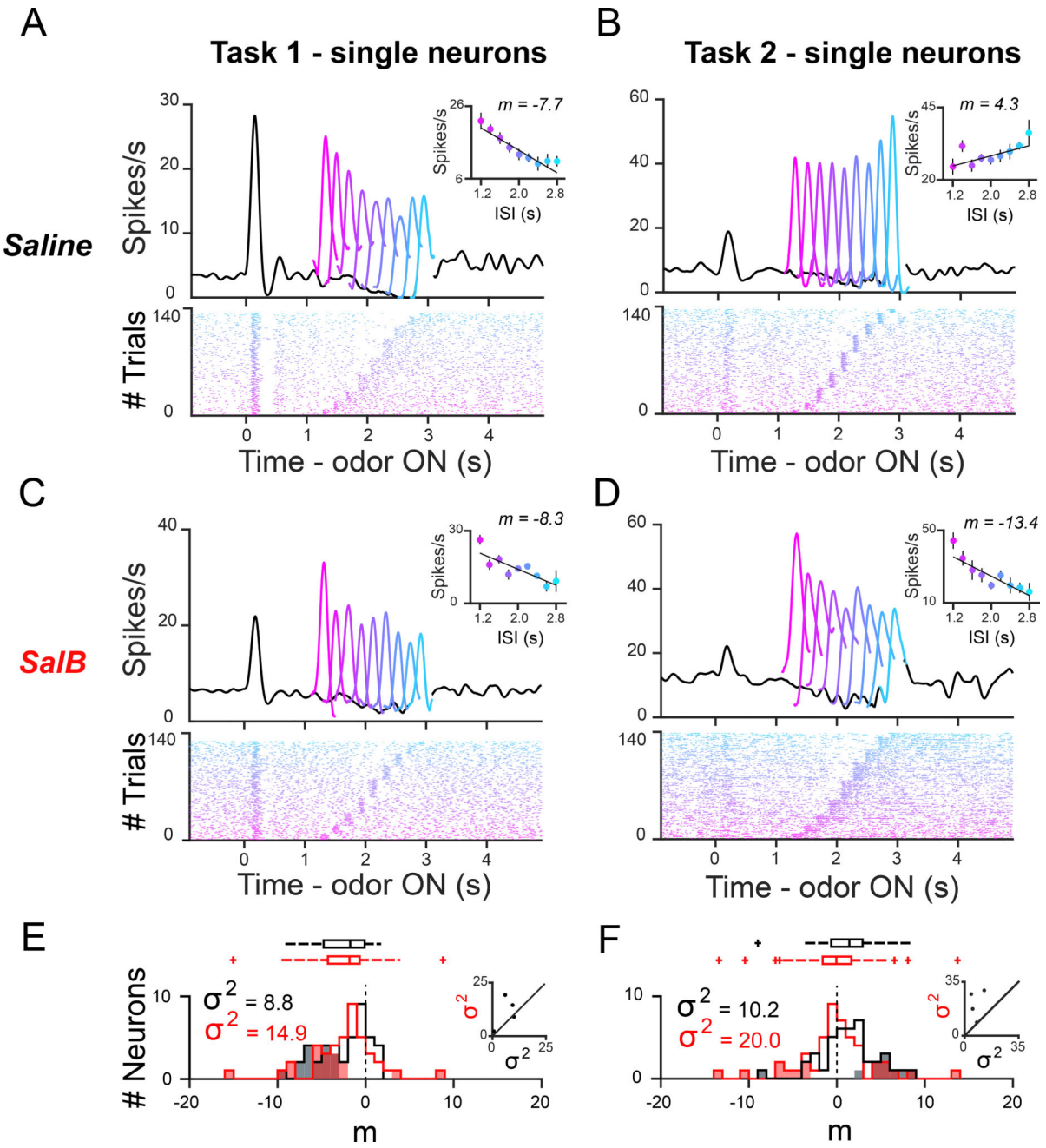


**Figure 5. Individual neurons recorded in Task 2 were more likely to show negative temporal modulation, following mPFC inactivation**

(A–D) PSTHs and raster plots for single neurons. A best-fit line was drawn through a plot relating the ISI to the post-reward firing rate for each odor A trial. The example 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 does not include 0. The distributions were significantly different in Task 2 ($median_{Saline} \neq median_{SalB}$, $z = 2.1$, $p = 0.04$,

Wilcoxon rank-sum test) but not in Task 1 ($median_{Saline} \neq median_{SalB}$, $z = 0.14$, p = 0.89, Wilcoxon rank-sum test). Inset: For individual KORD-expressing animals recorded in Task 1 (n = 4) and Task 2 (n = 4), variance of slopes in Saline versus SalB condition. The variance significantly increased in Task 2, in the SalB condition ( $\sigma^2_{Saline} \neq \sigma^2_{SalB}$, $F_{40,46} = 0.51$, p = 0.03, *F*-test for equality of two variances), but not in Task 1 ( $\sigma^2_{Saline} \neq \sigma^2_{SalB}$, $F_{41,41} = 0.59$, p = 0.10, *F*-test for equality of two variances).

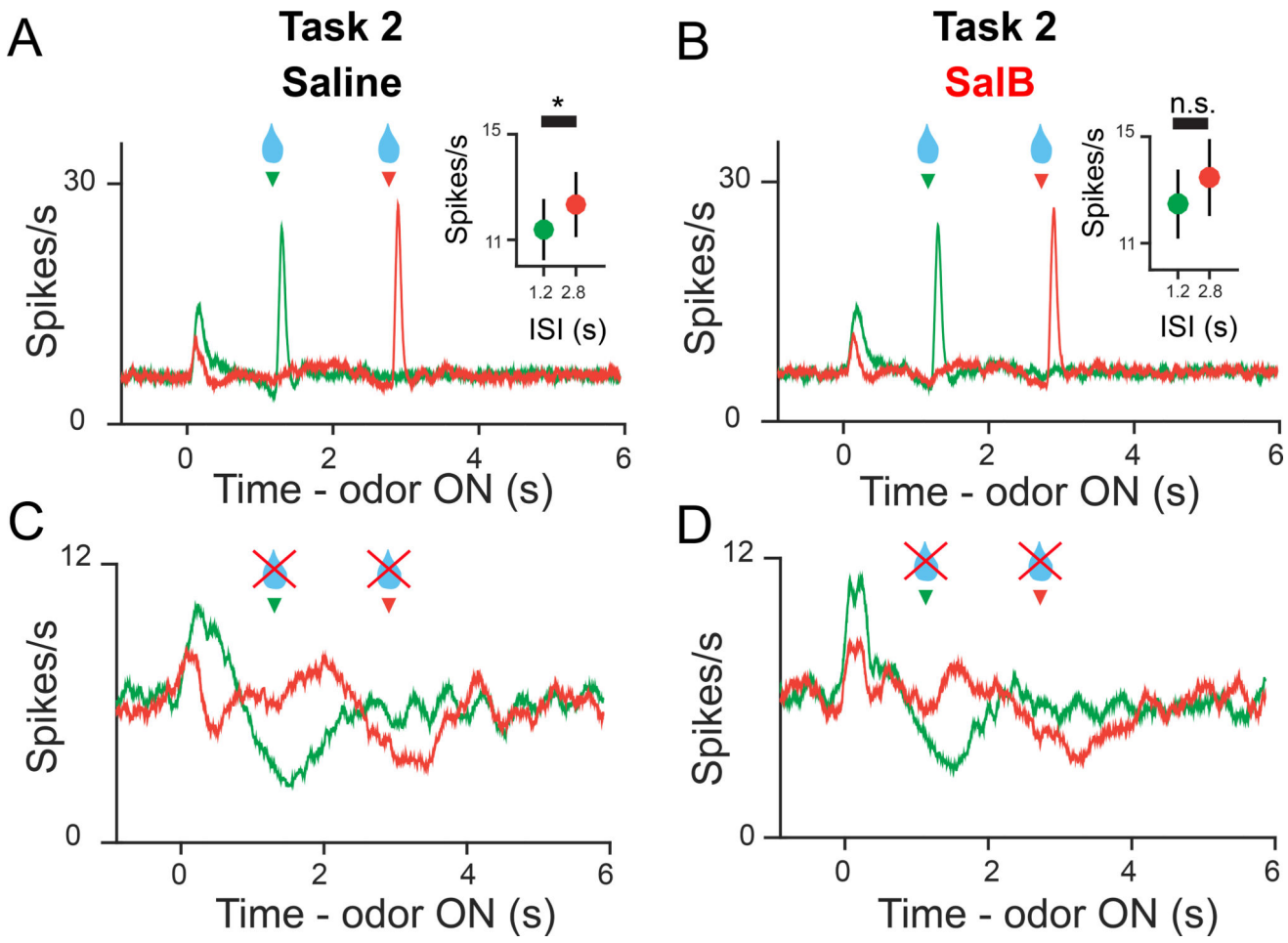


**Figure 6. mPFC inactivation spared timing-related aspects of dopamine signaling**

(A,B) PSTH of RPEs on rewarded Odor B (green) and C (orange) trials for all neurons recorded in Task 2, Saline condition (A) and SalB condition (B). Inset displays average post-reward RPEs for Odors B and C (mean ± SEM). Odor C post-reward RPEs were significantly larger than Odor B post-reward RPEs in the Saline condition (Absolute difference = 0.95Hz, $F_{1,40} = 5.8$, p = 0.02, 2-way ANOVA, factors: neuron, Odor) but not in the SalB condition (Absolute difference = 1.0Hz, $F_{1,40} = 2.6$, p = 0.11)

(C,D) PSTH of RPEs on omission Odor B and C trials for all neurons recorded in Task 2, Saline condition (C) and SalB condition (D). Omission 'dips' were significantly smaller than baseline in both Saline ($F_{1,40} = 37$, $p = 3.8 \times 10^{-7}$ (Odor B); $F_{1,40} = 14$, $p = 7.0 \times 10^{-4}$ (Odor C), 2-way ANOVA, factors: neuron, 0–1s following time of usual reward delivery versus 1-0s before odor onset), and SalB ($F_{1,40} = 13$, $p = 9.0 \times 10^{-4}$ (Odor B); $F_{1,40} = 12$, $p = 1.3 \times 10^{-3}$ (Odor C)).

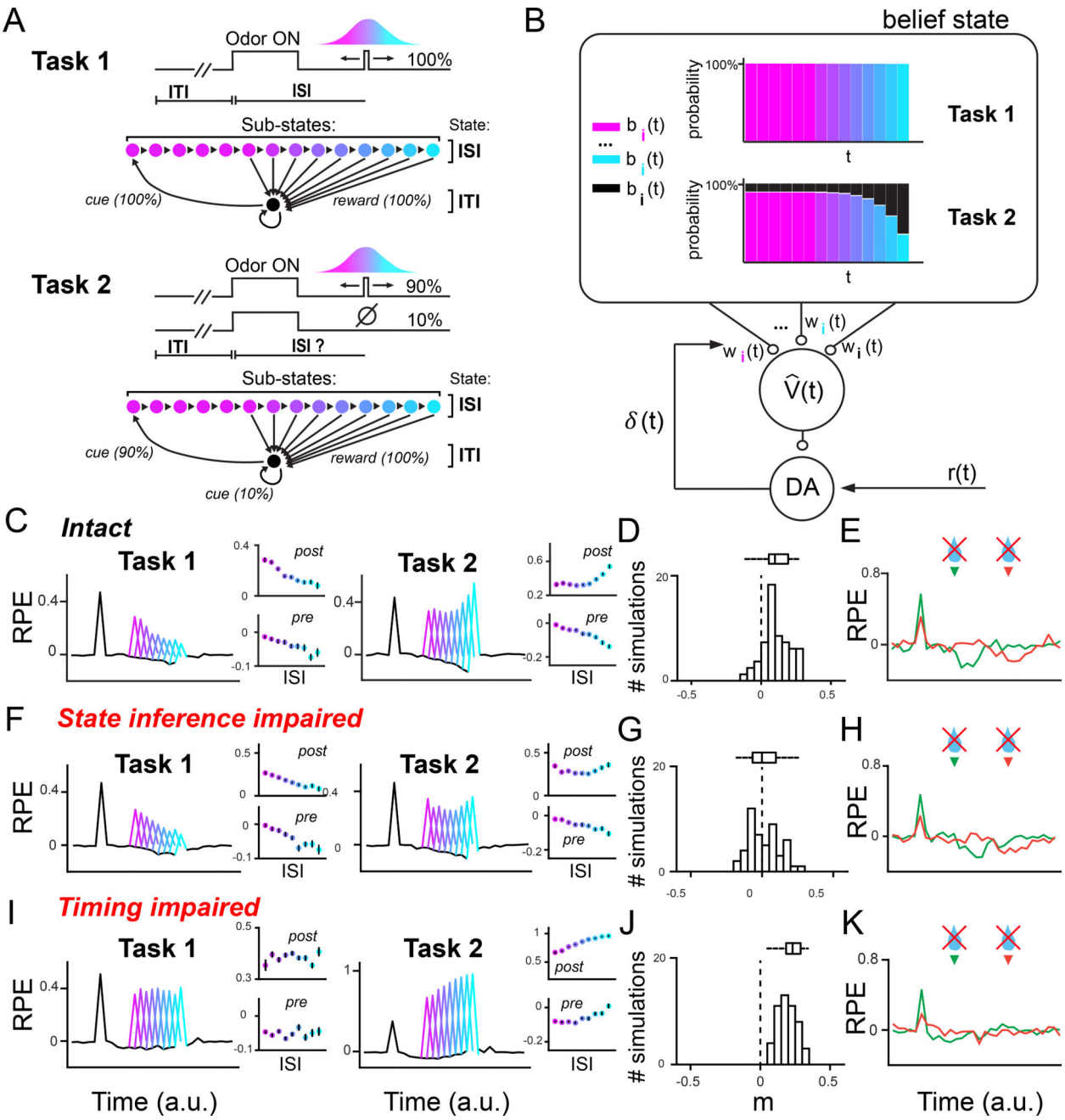


**Figure 7. Computational modeling implicates mPFC in computing the belief state**

(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.

(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 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. (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).

(D,G,J) Same analysis as shown in Figure 5, 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 that tends more towards negative values, similar to our data (Figure 5F).

(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.

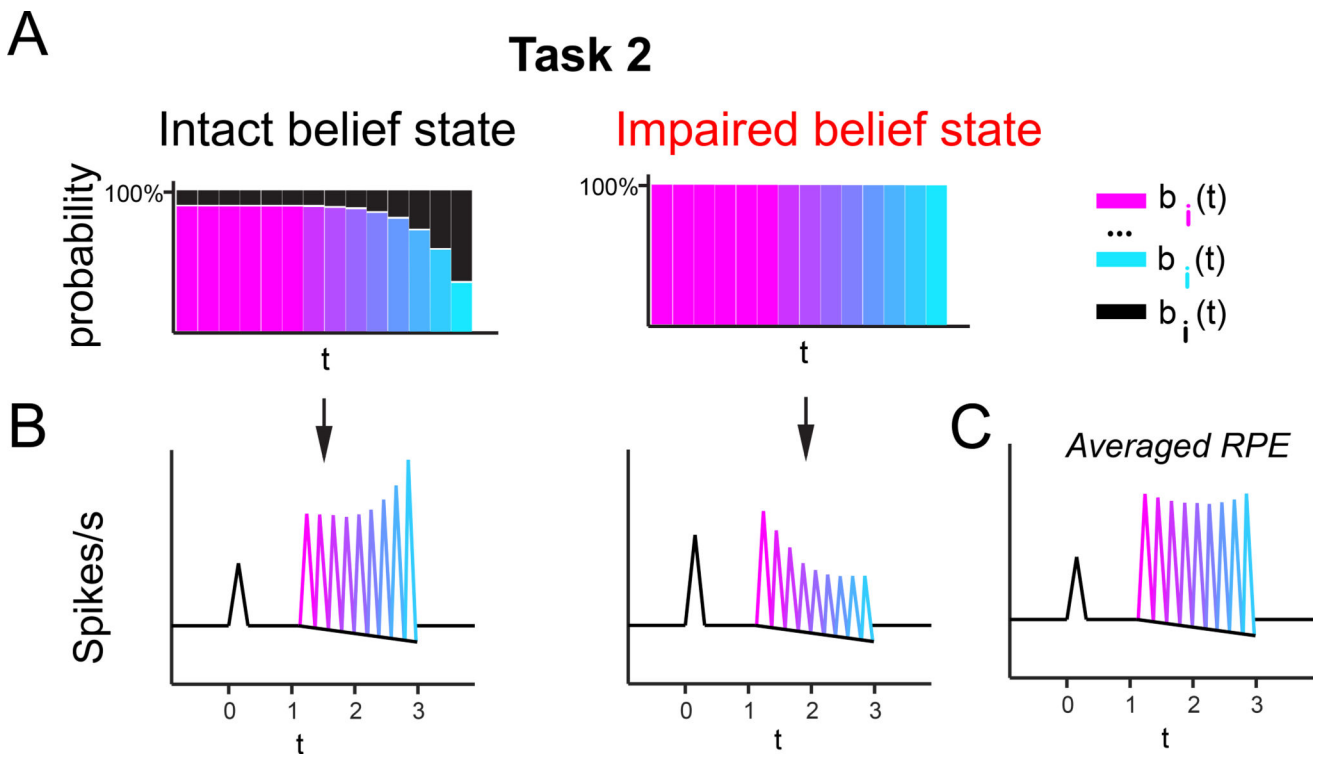


**Figure 8. Cartoon hypothesis of mPFC inactivation results, based on computational modeling**

(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 4D.

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Antibodies | | |
| Rabbit Anti-Tyrosine Hydroxylase | Millipore | AB_390204 |
| Bacterial and Virus Strains | | |
| AAV5-EF1a-DIO-hChR2(H134R)-EYFP | UNC Vector Core | N/A |
| AAV8/CamkII-KORD-IRED | UNC Vector Core | N/A |
| Experimental Models: Organisms/Strains | | |
| Mouse: Slc6a3$^{tm1(cre)Xz}$/J | The Jackson Laboratory | Jax #020080 |
| Software and Algorithms | | |
| MATLAB (version 2016a) | Mathworks | http://www.mathworks.com |
| LabView (version 2013) | National Instruments | http://www.ni.com |
| MClust software (version 4.3) | A. David Redish | http://redishlab.neuroscience.umn.edu/MClust/MClust.html |
| Other | | |
| Salvinorin B | Apple Pharms | N/A |
| Isosol (Isoflurane, USP) | Vedco | N/A |
| Ketoprofen (for analgesia) | Patterson Veterinary | Cat #07-803-7389 |
| Buprenorphine | Patterson Veterinary | Cat #07-850-2280 |
| Dexamethasone | Patterson Veterinary | Cat #07-808-8194 |
| LRS-0473 DPSS Laser System | LaserGlow Technologies | Cat #R471003FX |
| NI-DAQ card, PCI-e6251 | National Instruments | Cat #781048 |
| FT200EMT Custom Patch Cord Length: 2m End A: FC/PC End B: 1.25m (LC) Stainless Steel Ferrule Furcati | Thorlabs | N/A (Custom) |
| Digital Lynx 4SX | Neuralynx | N/A |
| Sandvik Kanthal HP Reid Precision Fine Tetrode Wire | Sandvik | Cat #PF000591 |