Supplemental Figures

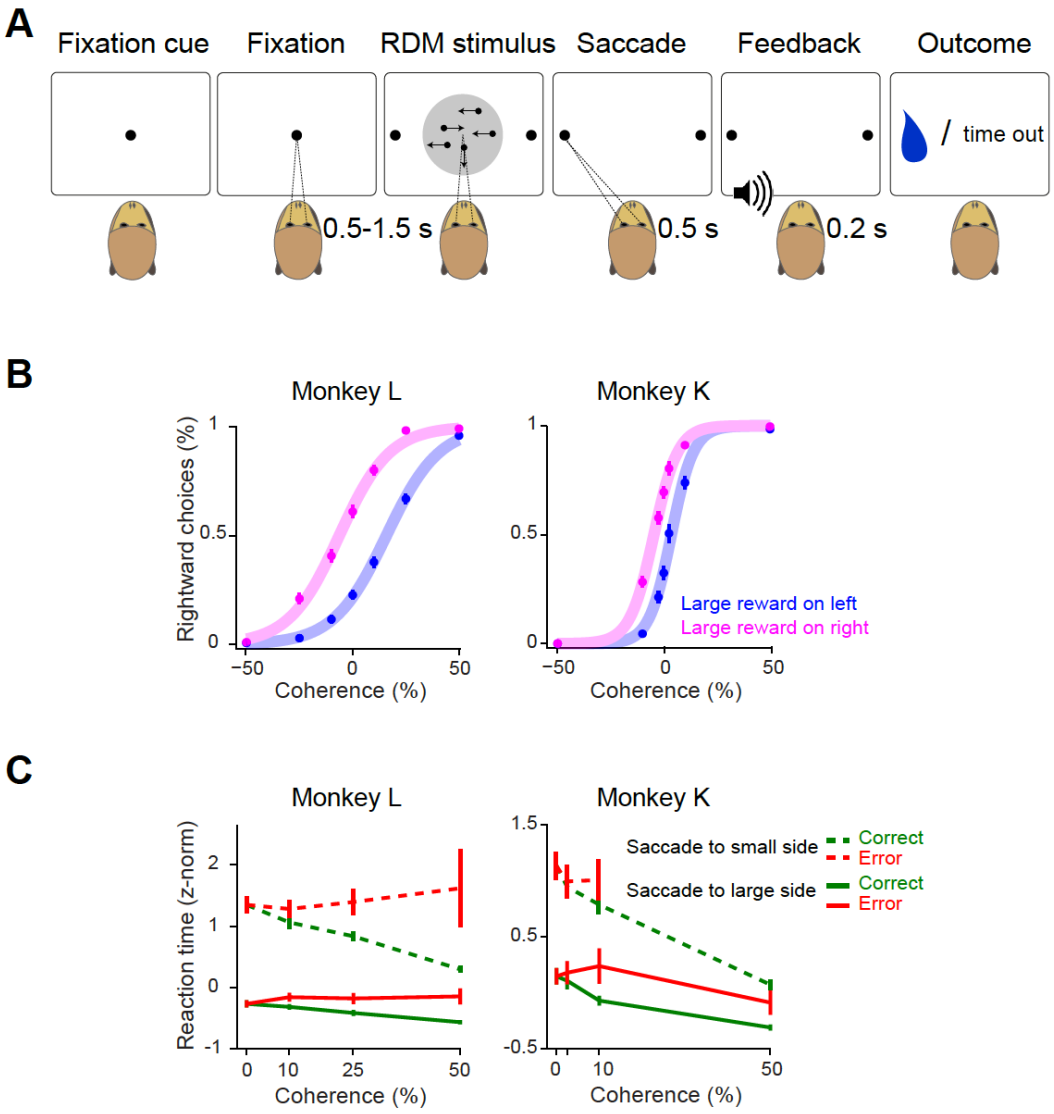

Figure S1. Monkeys' decisions reflect both stimulus difficulty and reward magnitude (Related to Figure 1).

(A) Monkeys' psychometric curves separated based on the response side to which the large reward magnitude was assigned. Animals could categorize easy random dot motion stimuli almost perfectly and were challenged with more difficult stimuli. Moreover, monkeys tended to respond in the direction associated with the large juice reward. Dots indicate data averaged across all testing sessions. Thick lines represent logistic fits to the data. Both animals showed significant bias towards the side with larger reward ($p < 0.05$, in both animals, permutation test). In all panels, error bars are s.e.m. across test sessions.

(B) Choice reaction time. The saccadic reaction times were z-normalized and separated based on motion coherence (its absolute value) and saccade direction (to the side associated with large or small reward). Monkeys showed faster reaction times when making saccade to the side associated with the larger reward (compare dashed lines with solid lines). Moreover, animals' reaction times were modulated by stimulus difficulty and decision outcome (i.e. correct or error) in a manner consistent with predictions of the TDRL model with belief state.

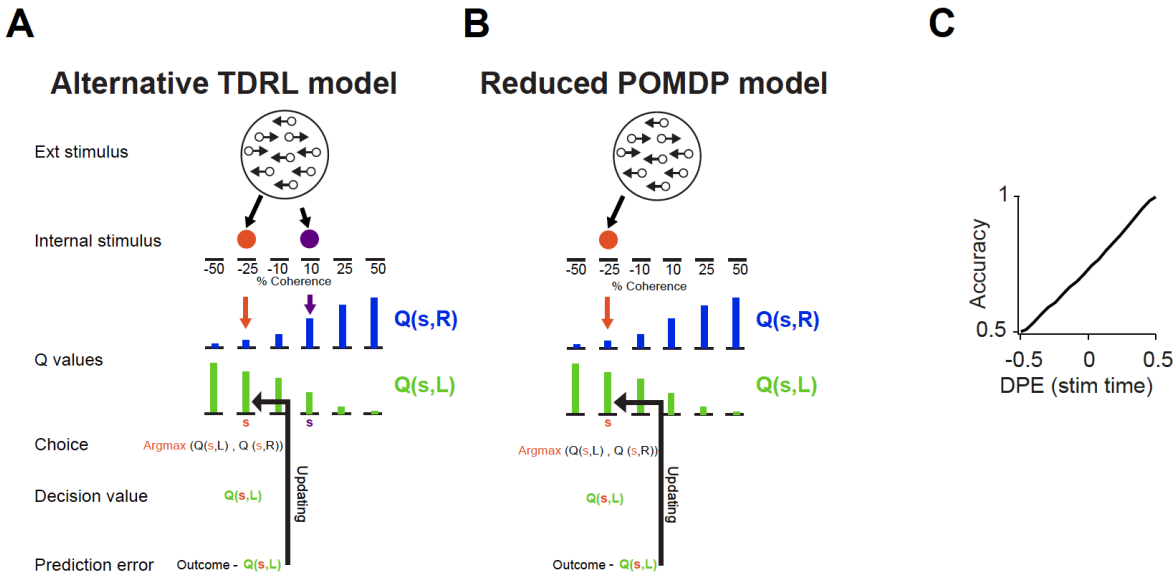

Figure S2. Schematic of the alternative model and the reduced POMDP model and additional predictions of the main TDRL model (Related to Figure 1).

(A) In this model, the decision making system assign one state, $\hat{S}_m$ (shown in orange), to the motion stimulus and makes the choice by comparing $Q(\hat{S}_m, L)$ and $Q(\hat{S}_m, R)$ $(a = \underset{A}{\operatorname{argmax}}\, Q(A))$. Since the dopamine system does not have direct access to the sensory evidence used for choice, it assigns another state, $\hat{S}_{m'}$ (shown in purple), to the motion stimulus, which could be identical to different from the one used for choice,$\hat{S}_m$. The larger Q-value $(Q(\hat{S}_{m'}, L)$or$Q(\hat{S}_{m'}, R))$ is used for prediction error computation.

The dopamine prediction error patterns of this model are shown in Figure 1F-H.

(B) Schematic of the reduced POMDP model. This model does not include a full belief state but uses the mean of the belief state to assign a single state $\hat{S}_m$ to the motion stimulus and perform choice by comparing $Q(\hat{S}_m, L)$ and $Q(\hat{S}_m, R)(a = \underset{A}{\operatorname{argmax}}\, Q(A))$. The prediction error patterns are similar to those of our full POMDP model (see Figure 1C-E). Such a reduced model could achieve what the full POMDP achieves in one trial, over many of trials.

(C) Decision accuracy of the TDRL model with full belief state as a function of decision value prediction errors (DPEs) at the time of stimulus.

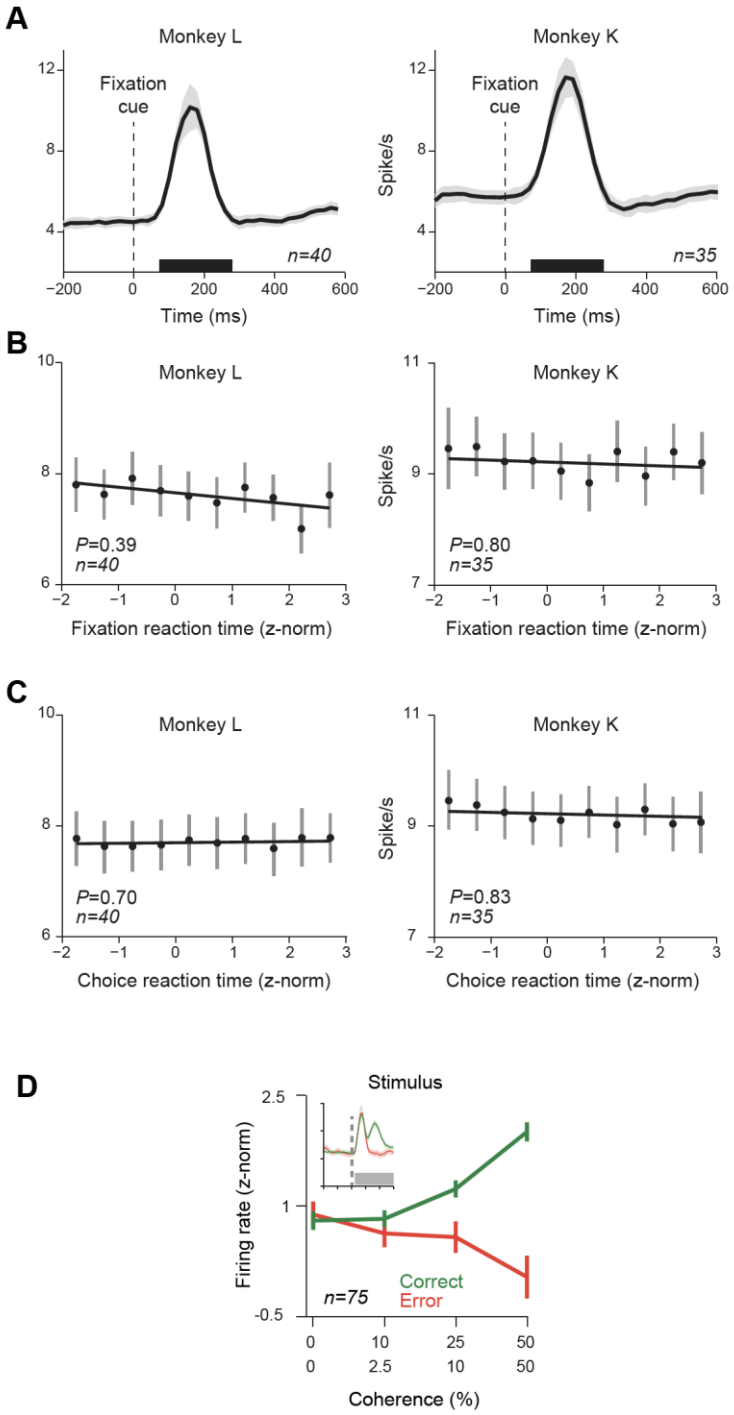

Figure S3. Dopamine responses to the fixation cue do not predict reaction times (Related to Figure 3).

(A) Dopamine population responses to the fixation cue. The black horizontal bar indicates the temporal window used for the analysis shown in (B) and (C).

(B) Dopamine responses to the fixation cue plotted as a function of z-scored fixation reaction time. In each panel of the figure, the line shows single linear regression on the population responses.

(C) Dopamine responses to the fixation spot as a function of z-scored choice reaction time.

(D) The population dopamine responses at the time of motion stimulus measured 60-600 ms after the stimulus onset.

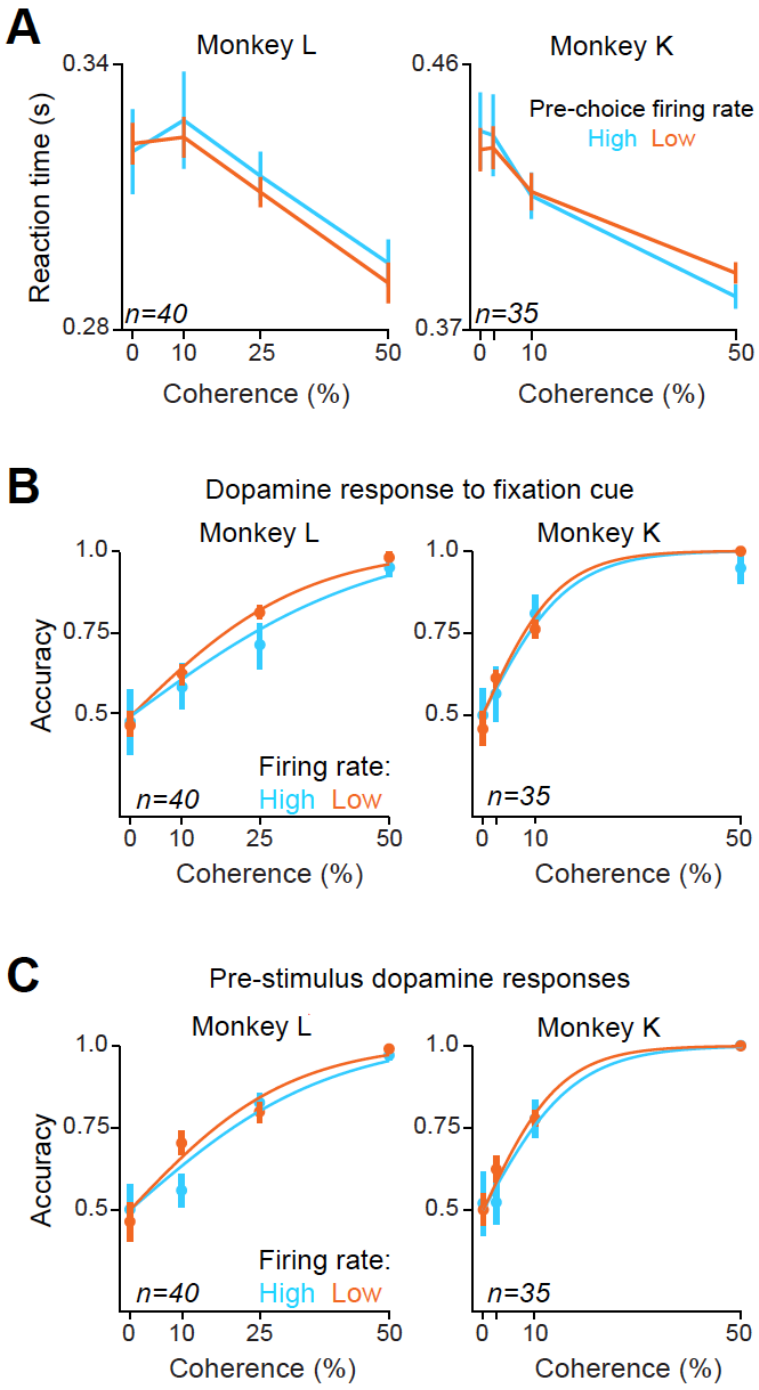

Figure S4. Pre-choice dopamine responses do not predict reaction times and fixation or pre-stimulus dopamine responses do not predict choice accuracy (Related to Figure 4).

(A) Animals' saccadic reaction times separated based on the pre-saccade dopamine responses (below and above 75th percentile, respectively).

(B) Choice accuracy as a function of dopamine responses to the fixation cue (below and above 75th percentile, respectively) computed separately for the two monkeys.

(B) Choice accuracy as a function of dopamine pre-stimulus tonic responses (below and above 75th percentile, respectively) computed separately for the two monkeys.

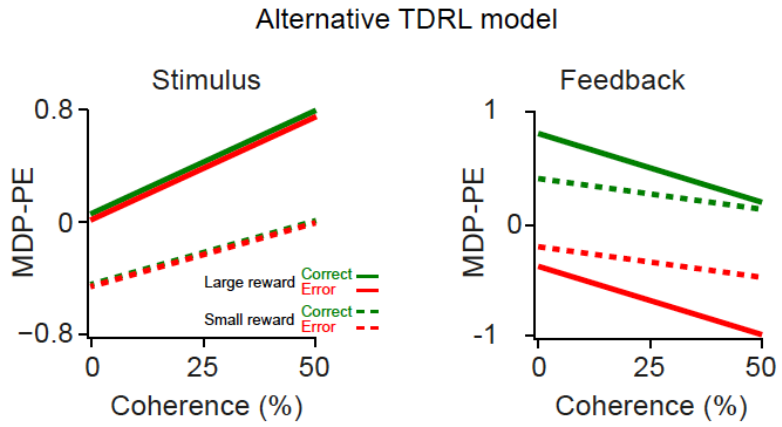

Figure S5. Prediction errors of the alternative TDRL model when all trials, regardless of reward size re included in the analysis (Related to Figure 5).

Supplemental Experimental Procedures

Temporal difference reinforcement learning models

Here we describe the basic features of the model implementation that were common among all model variants.

We simulated the sequence of behavioral events in each trial as *states, s*. For our task, these states are 'initial, 'fixation cue', 'motion stimulus', 'feedback and 'end', denoted as s_i, s_{fc}, s_m, s_{fb}, s_e. In each state, the agent performs an action, a, observes an outcome and transits to the next state, s'.

Apart from the 'motion stimulus' state, in which the agent learns which action (left or right) to take, in all other states the agent visits the subsequent state based on a pre-defined transition probability. This transition function indicates the probability that the agent visits the state s' from its current state s, as

$$p_{ss\prime} = p\{s_{t+1} = s' | s_t = s\} \quad \text{Eq. 1}$$

For instance, we set the probability of transition from the 'fixation cue' to the 'motion stimulus' to 0.99, meaning that in 99% of trials the agent visits 'motion stimulus' after the 'fixation cue' state. In the remaining 1% trials, after the 'fixation cue' the agent visits the 'trial end' state, resembling trials in which animals failed to fixate. These transition probabilities were set to reproduce animals' highly stable success in fixating on the fixation cue (~99% of trials) and were kept constant across all trials of the model run. For our model illustrations in Figure 1,2, 5 and Figure S5, we only include trials in which the agent reached 'motion stimulus' state.

The goal of the agent is to take actions that maximize the discounted cumulative reward, defined as:

$$R_t = \sum_{k=0}^{\infty} \gamma^k \, r_{t+k+1} \quad \text{Eq. 2}$$

where r_t is the immediate reward the agent receives in transitioning from s_{t-1} to s_t and γ is a discount factor that controls the degree to which immediate rewards are preferred to rewards achieved in subsequent state transitions.

When occupying state s, the state-action value, $Q(s, a)$, defines the expected cumulative reward when the agent occupies state s and takes action a:

$$Q(s, a) = E[\sum_{k=0}^{\infty} \gamma^k \, r_{t+k+1} | s_0 = s, a_0 = a] \quad \text{Eq. 3}$$

After the transition from s_t to s_{t+1}, the agent makes a comparison between the prior value prediction and current value estimate and computes a prediction error, defined as:

$$\delta_t = r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \quad \text{Eq. 4}$$

The agent uses the computed prediction error to update the action value estimates, using the following updating rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \delta_t \quad \text{Eq. 5}$$

where α is the learning rate. For our simulations we set $\alpha = 0.01$ and $\gamma = 1$ (i.e. no temporal discounting).

Behavioral task

The behavioral task has been described previously in detail [S1] and is outlined here briefly. Two male monkeys (Japanese macaques, weighing 7-9.5 kg) were rewarded in each trial for correct discrimination of the motion direction of a random dot motion stimulus. We used a set of random dot motion stimuli with two directions (right and left), and four coherence levels (0, 10, 25, and 50% for monkey L; 0, 2.5, 10, and 50% for monkey K). A trial started with the appearance of a fixation cue at the center of the monitor, followed by a dynamic random dot motion stimulus and two peripheral targets, after which the

monkey were free to make a saccade to one of two targets to indicate its choice. The random dot motion stimulus disappeared as soon as the monkey made an eye movement. Monkeys kept their gaze on the chosen target for 0.5 s and then received different auditory feedbacks for correct and error choices. If the monkey chose correctly, a high pitch feedback tone (1000 Hz, 0.2 s) was delivered, followed by a juice reward immediately after the tone offset. When the choice was incorrect, only a low pitch feedback tone (400 Hz, 0.2 s) was delivered, with an additional 5 s timeout as a penalty. Error trials were repeated to the animal and monkeys had near perfect performance in these repeat trials. Thus, it is more accurate to describe error trials as having delayed reward, rather than no reward. At the zero coherence level, motion direction was randomly assigned as either "rightward" in half of the trials or "leftward" in the other half. In each block of 126-168 trials, one direction of motion was associated with a large reward (0.38 ml), and the other was associated with a small reward (0.16 ml). The direction-reward contingency was fixed throughout a given block and reversed in the subsequent block. Animals could categorize easy (high motion coherence) stimuli almost perfectly but were challenged by more difficult stimuli (low motion coherence) and showed bias toward the direction associated with the large reward (Figure S1).

Analysis of the behavioral data

The behavioral data have been described in detail previously [S1]. We fitted the choice data to a logistic function (Figure S1A). For the analysis of choice reaction time (the interval between the onset of the random dot motion stimulus and the time that animal's saccade landed on one of the target) and fixation reaction time (the interval between the onset of the fixation cue and the time that animal's saccade landed on it), we normalized each trial's reaction time by computing session-by-session z-scored reaction times (Figure S1B and Figure S3B and C).

Localization and recording of dopamine neurons

Dopamine neuronal recording has been described in details previously [S1] and will be described here briefly. We estimated the location of the substantia nigra by proton density-enhanced magnetic resonance (MR) images. We placed a round recording chamber (Crist Instrument) on the skull with dental cement so that the center of the recording chamber targeted the substantia nigra pars compacta. Recordings were made using an epoxy-coated tungsten electrode (shank diameter, 0.25 mm, 0.5–1.5 MΩ measured at 1000 Hz (FHC). Dopamine neurons were identified according to their low tonic irregular spontaneous firing rates (<10 Hz), relatively long duration of action potentials (>1.5 ms), and transient responses to unexpected reward delivery.

Analysis of the neuronal data

The temporal windows used for the analysis of the neuronal data are shown in Figure 3, 4 and Figure S3 (post fixation cue: 80-280 ms, pre random dot motion stimulus (for tonic dopamine response): -500−0 ms, post random dot motion stimulus: 220−500ms, pre saccade: -300−0 ms, post feedback tone: 80−330 ms). Because dopamine neurons showed qualitatively similar responses in the present study, the time windows specified above were applied to all recorded neurons (apart from minimal modifications on the analysis time window used for illustrated example neurons, as shown with gray horizontal bars in Figure 3 and 4). We used raw neuronal firing rates for all our analysis, apart from the analysis shown in Figure 3D and 5B in which we z-scored normalized the activity of each neuron.

To quantify the time course of dopamine responses in the correct and error trials, we used sliding window receiver operating curve (ROC) analyses (sliding window of 250 ms shifted in 10 ms steps) aligned to different task events. We used the area under constructed ROC curve (AUC) as the index indicating differential neuronal activity in correct and error trails (AUCs close to 1 indicate larger dopamine responses in the correct trial compared to the error trials and AUCs close to 0 correspond to smaller neuronal responses in the correct trials compared to the error trials). To assess the statistical significance of computed AUCs, we used a permutation test (with 200,000 resamples) and determined the first instance that the AUC reached statistical significance during each trial by finding the time epoch that the permutation test indicated statistical significance ($P < 0.001$) in three consecutive time steps. We also used AUC measures to quantify neuronal response difference in a fixed time window after task events (as defined above) in correct/error trials as well as small/large reward trials (Figure 5C and D) and examined their statistical significance using permutation test, $P < 0.01$.

Supplemental References

S1. Nomoto, K., Schultz, W., Watanabe, T., and Sakagami, M. (2010). Temporally Extended Dopamine Responses to Perceptually Demanding Reward-Predictive Stimuli. J Neurosci *30*, 10692-10702.