Supporting Information

Sarno et al. 10.1073/pnas.1712479114

SI Materials and Methods

Detection Task. Two monkeys were trained to detect a vibrotactile stimulus of variable amplitude applied to one of each monkey's fingertips (6). Stimulus-present trials were randomly interleaved with an equal number of stimulus-absent trials. Stimuli were delivered to the skin of the distal segment of one digit of the restrained hand, via a computer-controlled stimulator (2-mm round tip; BME Systems). Initial probe indentation was 500 μm. Vibrotactile stimuli consisted of trains of 20-Hz mechanical sinusoids with nine different amplitudes between 2.3 μm and 34.6 μm. Crucially, some of the amplitudes were very weak and consequently difficult to detect. Animals were rewarded with a drop of liquid for correct behavioral responses (correct detections in stimulus-present trials and CRs in stimulus-absent trials) and received no reward otherwise (miss trials and FA trials).

Recordings. Data for this analysis were obtained from an earlier study (27). Recordings were obtained with quartz-coated platinum–tungsten microelectrodes (2–3 MΩ; Thomas Recording) inserted through a recording chamber located over the central sulcus, parallel to the midline. Midbrain DA neurons were identified on the basis of their characteristic regular and low tonic firing rates (1–10 spikes per second) and by their long extracellular spike potential (2.4 ms ± 0.4 SD). Among the 69 neurons analyzed in the previous work we selected a group of 23 cells (monkey A, $n = 9$; monkey B, $n = 14$). The selected group of cells corresponded to those neurons whose response to the reward delivery did not violate a RL principle: They showed a positive phasic activation or lack of response in correct trials (hit and CR trials) while the activity paused or remained at the baseline level when the reward was omitted (miss and FA trials). A similar criterion has been adopted in many electrophysiological studies of midbrain DA neurons (23, 33). The recorded sites of the selected neurons differed from the discarded ones only in their depth (the antero-posterior and medio-lateral coordinates were kept constant). The median depth of the 23 selected neurons was 362 μm above the median of the other 46 neurons. A two-sample t test between the depths of the two groups of neurons showed that their difference was at the margin of statistical significance ($P = 0.055$).

Data Analysis. For each neuron, we computed the firing rate as a function of time, using 300-ms sliding windows displaced every 50 ms (Fig. 1*B*). Responses to the stimulus (Fig. 1*C* and in Fig. 2*A*, *Right*) were measured in a 500-ms window centered 350 ms after the stimulus onset and were standardized with respect to a prestimulation window (of 500 ms centered 700 ms before the stimulus presentation). Responses to the go instruction (Fig. 3*A*) were measured in a 250-ms window centered 170 ms after the instruction and were standardized with respect to a precue window (of 250 ms centered 500 ms before the cue presentation). Responses to the reward delivery were measured in a 400-ms window centered 350 ms after the PB and were standardized with respect to a precue window of 200 ms centered 200 ms before the PB (Fig. 3*B*). The activity outside the PSW was calculated in two 1-s windows before the start and after the end of the PSW (from 500 ms to 1.5 s after the KD event and from 3.7 s to 4.7 s after that event). The mean activity during and outside the PSW was standardized with respect to a 500-ms window centered 1 s after the KD event (Fig. 2*B*). To determine the statistical significance of the computed AUROCs in Fig. S5, we used a permutation test with 10,000 resamples (significance was assessed when the permutation test indicated $P < 0.01$).

Model. The model relies on two modules: a Bayesian module and a RL module.

Bayesian module. This module uses noisy observations to estimate a posterior probability (belief) about the current state of the external world, s_t. More specifically, it calculates the belief $b_{sp}(t)$ about the presence of the (ambiguous) vibrotactile stimulus,

$$b_{sp}(t) = P(s_t = s_p | X_{1:t}), \quad \textbf{[S1]}$$

where $X_{1:t}$ is the entire history of observations up to time t. In what follows we describe the detailed equations used by the Bayesian module. This module represented some high-level cortical areas receiving inputs from sensory areas. We referred to these inputs as observations x_t and interpreted them as Poisson trains with firing rates $\lambda_i (i = 0, \ldots, N_a)$.

Each λ_i corresponded either to the absence of a vibrotactile stimulus ($i = 0$) or to the application of that stimulation with one of the $N_a = 9$ possible values of its amplitude during the time step t. Each of the 10 mean firing rates corresponded to a state i of the world. In each time step t the module computed a posterior probability (belief) $b_t(i)$ about the hidden state of the world, using the entire history of observations up to time t:

$$b_t(i) = P(\lambda_t = \lambda_i | X_{1:t}) 0. \quad \textbf{[S2]}$$

The beliefs about the absence and the presence of the stimulus corresponded, respectively, to

$$\begin{aligned} b_t(sa) &= P(\lambda_t = \lambda_0 | X_{1:t}) \\ b_t(sp) &= \sum_{i \neq 0} P(\lambda_t = \lambda_i | X_{1:t}). \end{aligned} \quad \textbf{[S3]}$$

Due to the complex temporal structure of the task, evaluating the $b_t(i)$ required estimating the joint posteriors $\tilde{b}_t(i,n)$ on the value of the firing rate of the input (λ_i) and the time n elapsed since the environment underwent a change to the state i. We therefore computed the belief over λ_t by marginalizing:

$$b_t(i) = \sum_n P(\lambda_t = \lambda_i, l_t = n | X_{1:t}) = \sum_n \tilde{b}_t(i,n). \quad \textbf{[S4]}$$

We separated the last part of the history, i.e., the last observation x_t, and calculated each belief recursively over time, using Bayes' rule,

$$\begin{aligned} \tilde{b}_t(i,n) &= P(\lambda_t = \lambda_i, l_t = n | X_{1:t-1}, x_t) \\ &= k . P(x_t | \lambda_t = \lambda_i) \sum_n P(\lambda_t = \lambda_i, l_t = n | X_{1:t-1}), \end{aligned} \quad \textbf{[S5]}$$

where $k = P(x_t | X_{1:t-1})$ is a normalization constant. The second term in Eq. **S5** was simplified using the Markov assumption and the fact that x_t did not depend on the length l_t (it depends only on the firing rate at the current time, λ_t). This term in Eq. **S5** represented the observation probability (*Observation probabilities*). The last term in Eq. **S5** could be rewritten as follows:

$$P(\lambda_t=\lambda_i,l_t=n|X_{1:t-1})=\sum_{j,m}\Big[P\big(\lambda_t=\lambda_i,l_t=n|\lambda_{t-1}=\lambda_j,l_{t-1}=m,X_{1:t-1}\big)\times P\big(\lambda_{t-1}=\lambda_j,l_{t-1}=m|X_{1:t-1}\big)\Big]$$
$$=\sum_{j,m}\Big[P\big(\lambda_t=\lambda_i|\lambda_{t-1}=\lambda_j,l_{t-1}=m,l_t=n,X_{1:t-1}\big)\times P\big(l_t=n|\lambda_{t-1}=\lambda_j,l_{t-1}=m,X_{1:t-1}\big)\times\tilde{b}_{t-1}(j,m)\Big]. \quad \textbf{[S6]}$$

Eq. **S5** together with Eq. **S6** represented a recursive relationship for the joint posteriors $\tilde{b}_t(i,n)$. Evaluating them required the knowledge of the change-point prior $CPP(l_t,l_{t-1},\lambda_{t-1},X_{1:t-1},t-1)=P(l_t=n|l_{t-1},\lambda_{t-1},X_{1:t-1})$ and of the transition probability $P(\lambda_t=\lambda_i|\lambda_{t-1}=\lambda_j,l_{t-1}=m,l_t=n,X_{1:t-1})$.

The change-point prior resulted independent from the history $X_{1:t-1}$ and, taking into account that the run length either increased by one after each time step or became zero at a change point, the *CPP* could be expressed as

$$CPP(n,m,\lambda_j,t-1)=\begin{cases}1-h(\lambda_j,m,t-1) & \text{if } n=m+1\\ h(\lambda_j,m,t-1) & \text{if } n=0\\ 0 & \text{otherwise.}\end{cases} \quad \textbf{[S7]}$$

The function $h(\lambda_{t-1},l_{t-1},t-1)$ represented the hazard rate, i.e., the probability that a change point occurred at time $t-1$ given that the state of the world was λ_{t-1} for exactly l_{t-1} time steps. It could be defined accordingly to the task structure (*Hazard rate*). The third term of Eq. **S6**, i.e., the transition probability, could be written as

$$P(\lambda_t=\lambda_i|\lambda_{t-1}=\lambda_j,l_{t-1}=m,l_t=n)=\begin{cases}\delta_{ij} & \text{if } n=m+1\\ T_{ij} & \text{if } n=0\\ 0 & \text{otherwise,}\end{cases} \quad \textbf{[S8]}$$

where δ_{ij} represented the Kronecker delta and we introduced the matrix $T_{ij}=P(\lambda_t=\lambda_i|\lambda_{t-1}=\lambda_j,l_t=0)$ representing the transition probability conditioned to the occurrence of a change point (*Transition probabilities*). Using Eqs. **S7** and **S8** we could rewrite Eq. **S5** as

$$\tilde{b}_t(i,0)\propto\sum_{j\neq i}\sum_m T_{ij}h(\lambda_j,m,t-1)\tilde{b}_{t-1}(j,m)$$
$$\tilde{b}_t(i,n\neq 0)\propto[1-h(\lambda_i,n-1,t-1)]\tilde{b}_{t-1}(i,n-1). \quad \textbf{[S9]}$$

The equations above completely described the temporal evolution of the $\tilde{b}_t(i,n)$ once the hazard rate h and the transition probability matrix T_{ij} were defined.

Transition probabilities. Given that the transition matrix T_{ij} was conditioned to the occurrence of a change point, we needed only to define the quantities $T_{i\neq 0,sa}$ and $T_{sp,i\neq 0}$. These probabilities were independent from the particular value of the firing rate λ_i in the stimulus-present condition. We obtained that $T_{i\neq 0,sa}=1/9$ (because all of the nine amplitude values were equally probable) and $T_{sp,i\neq 0}=1$ (because the delay period always followed the stimulation).

Hazard rate. As for the transition matrix, the hazard rate for the stimulus-present condition was independent from the particular value of the firing rate λ_i. The hazard rate depended only on the time $t-1$, on the duration of an epoch before the transition, l_{t-1}, and on the state corresponding to that epoch, λ_{t-1}.

In the stimulus-absent condition this function took a value different from zero only during the PSWs and depended on the epoch length λ_{t-1} and on the time $t-1$ (because transitions were not allowed during the delay period). We defined it as

$$h(\lambda_{j-1}=\lambda_0,l_{t-1}=m,t-1)=\begin{cases}h_{sa}(m) & \text{if } m=t-1\\ 0 & \text{otherwise.}\end{cases} \quad \textbf{[S10]}$$

In the stimulus-present condition, given the task, the hazard rate depended only on the duration of the epoch before the transition and was defined as

$$h(\lambda_{j-1}\neq\lambda_0,l_{t-1}=m,t-1)=h_{sp}(m). \quad \textbf{[S11]}$$

The exact form of the functions $h_{sa}(m)$ and $h_{sp}(m)$ depended on the task temporal structure. If the interval timing mechanism was perfect, the function $h_{sa}(m)$ would represent the hazard rate corresponding to a uniform probability density function while $h_{sp}(m)$ would represent the hazard rate corresponding to a fixed duration interval lasting the stimulation period.

Nevertheless, these definitions ignored the fact that animals' interval timing processes did not take place with infinite accuracy (the accuracy of temporal estimation is supposed to be constrained by Weber's law). Following ref. 44 we calculated a "subjective" hazard function (based on the assumption of timing scalar noise) and used these subjective hazards to perform the inference. The value of the Weber fraction for time estimation used in the simulations was $\phi=0.18$.

Observation probabilities. The last step to implement Eq. **S5** was to define the quantities $P(x_t|\lambda_t)$. We considered that the observation x_t represented the number of spikes produced in a sensory area on a given time step and it was generated from a Poisson distribution with mean λ_t. The parameter λ represented the mean firing rate of a sensory area. Depending on the presence of the stimulus and on the amplitude value, the parameter λ_t could take the value λ_0, in stimulus-absent conditions, and the value λ_i, with $i\neq 0$, when a stimulus with amplitude i is presented. Therefore, we defined the observation x_t as follows:

$$x_t=\begin{cases}Poisson(\lambda_0) & \textit{if the stimulus is absent}\\ Poisson(\lambda_i) & \textit{if the stimulus is present with amplitude } i.\end{cases} \quad \textbf{[S12]}$$

We defined the probability to obtain the observation x_t given a mean firing rate λ_i at time t as

$$P(x_t|\lambda_i)=P_{poisson}(x_t|\lambda_i), \quad \textbf{[S13]}$$

where $P_{poisson}(x|\lambda)$ indicated the probability to obtain the observation x given a Poisson process with mean λ. The 10 values of the parameters λ_i were obtained from previously recorded data of the same experiment (6) and corresponded to the mean firing rates of a sensory area in the 10 different conditions. Their values, ordered according to increasing values of the amplitude of the stimulus, were 15 Hz, 15.2 Hz, 15.5 Hz, 16 Hz, 17 Hz, 20 Hz, 23 Hz, 27 Hz, 35 Hz, and 40 Hz.

Belief equations. Using Eq. **S9** the posterior probability $b_t(i)$ of being in the state i could be expressed as

$$b_t(i)=\sum_n\tilde{b}_t(i,n)\propto\sum_{j\neq i}\sum_m T_{ij}h_j(m,t-1)\tilde{b}_{t-1}(j,m)+\sum_{n\neq 0}[1-h_i(n-1,t-1)]\tilde{b}_{t-1}(i,n-1). \quad \textbf{[S14]}$$

For the stimulus-absent state the above equation took the form

$$b_t(sa) \propto \sum_{j\neq 0}\sum_{m} T_{sa,j}h_j(m,t-1)\tilde{b}_{t-1}(j,m) + \sum_{n\neq 0}[1-h_{sa}(n-1,t-1)]\tilde{b}_{t-1}(sa,n-1). \quad \textbf{[S15]}$$

Using the fact that $\tilde{b}_t(sp,m) = \sum_{j\neq 0}\tilde{b}_t(j,m)$ and the considerations about the hazard rate and the transition probabilities made in the previous sections, we obtained that

$$b_t(sa) = k\cdot P(x_t|sa)\left[\sum_m h_{sp}(m)\tilde{b}_{t-1}(sp,m) + \sum_{n\neq t}\tilde{b}_{t-1}(sa,n-1) + [1-h_{sa}(l_{t-1}=t-1)]\tilde{b}_{t-1}(sa,t-1)\right]. \quad \textbf{[S16]}$$

The first two terms of Eq. **S16** represented the probability of the delay interval while the last term corresponded to the probability of remaining within the prestimulus interval. Using Eq. **S9** we could define $b_t(\lambda_i \neq \lambda_0)$ for each of the nine amplitudes (with $\lambda_i \neq \lambda_0$) as follows:

$$b_t(i\neq 0) = k\cdot P(x_t|\lambda_i)\left[\sum_m T_{i\neq 0,sa}h_{sa}(t-1)\tilde{b}_{t-1}(sa,t-1) + \sum_{n>0}[1-h_{sp}(n-1)]\tilde{b}_{t-1}(i,n-1)\right]. \quad \textbf{[S17]}$$

Taking into account that $b_t(sp) = \sum_i b_t(i\neq 0)$ and the considerations about the transition probabilities and the hazard rate, we obtained

$$b_t(sp) = k\cdot\left[1/9\sum_i P(x_t|\lambda_i)\right]\sum_m h_{sa}(t-1)\tilde{b}_{t-1}(sa,t-1) + k\cdot\left[\sum_{n>0}[1-h_{sp}(n-1)]\sum_i P(x_t|\lambda_i)\tilde{b}_{t-1}(i,n-1)\right]. \quad \textbf{[S18]}$$

The former term in the above equation represented the probability of stimulus onset while the latter was the probability of remaining in a stimulus-present state condition before the stimulus offset (but after the onset of the vibration).

The stimulus was detected by the Bayesian module when the belief about its presence exceeded the belief about its absence:

$$b_t(sp) > b_t(sa) \Rightarrow \textit{stimulus detected}. \quad \textbf{[S19]}$$

The RL module. The latter module consists of a standard RL architecture known as actor/critic (18). We consider a total of six events: the vibrotactile stimulus, the start and go signals, and the response movements of the animal (KD and the two PBs indicating yes/no responses).

The physical salience function of event i is represented by the ith component of the vector. With the exception of the vibrotactile stimulus, the component $\boldsymbol{e}(t)$ takes value one at the onset of the event i and zero otherwise. The component $e_v(t)$ corresponding to the vibrotactile stimulus is activated when the Bayesian module detects it. In this case we set $e_v(t_d) = b_{sp}(t_d)$ (with t_d denoting the time of the detection).

The onset of the salience function $e_i(t)$ at time t^i_{on} activates a temporal representation $\boldsymbol{x}_i(t)$ of the event i. Since the stimulus has to be represented during a long delay period, we have used a temporal representation with optimal accuracy given a fixed number of resources (53). This is defined as a set of N functions $T_{im}(t)$ $(m = 1, \ldots, N)$, each representing the event (a pulse of one time step duration) around time τ_m after its detection. We assume that the resolution of these functions decreases with τ_m and that the times τ_m are distributed uniformly on a logarithmic timescale (from a minimum value $\tau_{min} = 0.1$ s to a maximum value $\tau_{max} = 10$ s). This leads to a scale-invariant representation of the event i. An explicit mathematical realization is (53)

$$T_{im}(t) \equiv T_i(t - t^i_{on}, \tau_m) = \frac{1}{|\tau_m|}C(k)\int_{d_i(t)}^{a_i(t)}\left(\frac{\tau'}{\tau_m}\right)^k e^{-k\frac{\tau'}{\tau_m}}d\tau', \quad \textbf{[S20]}$$

where $C(k) = k^{k+1}/k!$, $a_i(t) = t^i_{on} - t$, $d_i(t) = t^i_{on} + dt - t$, and dt is the duration of the original pulse (alternatively, Eq. **S19** could be expressed as a convolution of an alpha function with a pulse). The parameter k controls the smear in the representation (the larger k is, the more accurate the representation). The temporal representation $\boldsymbol{x}_i(t) = \{x_{i1}(t), x_{i2}(t), \ldots, x_{N1}(t)\}$ is taken equal to the functions in Eq. **S19** multiplied by the physical salience function of the event i:

$$x_i(t) = e_i(t^i_{on})T_i(t). \quad \textbf{[S21]}$$

The reward predicted by the event i is expressed as

$$P_i(t) = \sum_{m=1}^{N} x_{im}(t)w_{im}. \quad \textbf{[S22]}$$

The total predicted reward at time t, $V(t)$ is given by

$$V(t) = \sum_i P_i(t). \quad \textbf{[S23]}$$

Following ref. 42, we suppose that the occurrence of an event i with reward prediction higher than the total reward prediction at the previous time disrupts earlier events representations:

$$P_i(t^i_{on}) > \frac{V(t^i_{on}-1)}{\gamma} \Rightarrow x_{jm} = 0\,, \quad j\neq i. \quad \textbf{[S24]}$$

The DA signal is assumed to be represented by the RPE. However, DA neurons show an asymmetrical activity due to their low baseline firing rate. This asymmetry is taken into account by introducing a rectification threshold $\psi > 0$ for the RPE,

$$\delta(t) = \begin{cases} r(t) + TD(t) & \textit{if } r(t) + TD(t) > \psi \\ -\psi & \textit{otherwise}, \end{cases} \quad \textbf{[S25]}$$

where $TD(t) = \gamma V(t) - V(t-1)$ and $r(t)$ takes the value of R if the reward occurs at time t and 0 otherwise. The ratio between the value of ψ and the scalar reward value R determined the degree of asymmetry in the error signal (the asymmetry increases if the ratio decreases). The weights w_{im} in Eq. **S21** are adapted during learning as

$$\Delta w_{im} = \begin{cases} \eta_c^+ x_{im}\delta(t) & \textit{if } \delta(t) > 0 \\ \eta_c^- x_{im}\delta(t) & \textit{if } \delta(t) < 0, \end{cases} \quad \textbf{[S26]}$$

where η_c^+ indicates the learning rate for acquisition and η_c^- is the learning rate in extinction.

The input to the actor component is a vector trace $\bar{\boldsymbol{e}}(t)$ whose components $\bar{e}_i$ are defined as

$$\bar{e}_i(t) = e_i(t) + \rho \bar{e}_i(t-1), \qquad \textbf{[S27]}$$

where $\rho < 1$ is a decay parameter. The actor selects an action a_j only at the end of each trial, after the go cue. The possible actions are pressing one of the two buttons corresponding to yes/no decisions (the action of withholding movement is not allowed). The probability of choosing the action a_j for an input $\bar{e}(t)$ is given by a softmax distribution

$$P(a_j|\bar{e}(t)) = \frac{exp^{\frac{\sum_i \nu_{ij}\bar{e}_i}{\beta}}}{Z}, \qquad \textbf{[S28]}$$

where Z is the normalization constant and the parameter β governs the exploration/exploitation trade-off: As β approaches 0, action selection approaches a winner-take-all mode while larger values of β favor exploration. The weights ν_{ij} in Eq. **S27** are adapted only at the end of each trial when the reward is expected. Pressing of one of the two buttons occurs 0.3 s after the go cue. The reward is delivered 0.2 s after the movement. The weights ν_{ij} are adapted with the learning rule

$$\Delta\nu_{ij} = \begin{cases} \eta_a^+ \sum_t \bar{e}_i(t_r)\delta(t) & if\ j = \bar{j}\ , \delta(t) > 0 \\ \eta_a^- \sum_t \bar{e}_i(t_r)\delta(t) & if\ j = \bar{j}\ , \delta(t) < 0 \\ 0 & if\ j \neq \bar{j}, \end{cases} \qquad \textbf{[S29]}$$

where $\bar{j}$ denotes the selected action and t_r is the time when the reward is expected (i.e., five time steps after the go cue). The parameters η_a^+ and η_a^- correspond to the learning rate in acquisition and in extinction.

Model analysis. In all of the simulations we used a time bin dt = 100 ms (for a full list of parameters used in the model see Table S1). To compare the model results with the mean activity of DA neurons we transformed the simulated RPE $\delta(t)$ in an equivalent firing rate $[\delta(t)]_{equiv}$ as follows:

$$[\delta(t)]_{equiv} = baseline + F\delta(t). \qquad \textbf{[S30]}$$

The *baseline* representing the baseline activity of DA neurons during the trial was set to 5.1 Hz. The value of the scale factor F was chosen to obtain an equivalent prediction error $[\delta(t)]_{equiv}$ that matched the mean DA response at the start cue. Its value in all of the simulations was 27.5 Hz. Additionally, the signal $[\delta(t)]_{equiv}$ was filtered using a 300-ms sliding window displaced every 100 ms (a procedure equivalent to the one done to obtain the firing rate of DA neurons as a function of time). Responses to the stimulus (in Fig. 6*B*) were calculated, averaging the signal $[\delta(t)]_{equiv}$ over a 300-ms window centered 100 ms after the stimulus onset. Responses to the go instruction and to the reward delivery were calculated, averaging the signal $[\delta(t)]_{equiv}$ over a 300-ms window centered, respectively, 100 ms after the go cue and after the reward delivery (Fig. 6*A*).

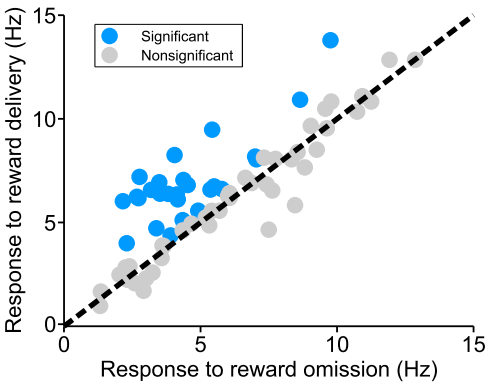

Fig. S1. Selection of midbrain neurons. The neurons used for the study (n = 23) corresponded to those cells whose responses to the reward delivery in correct trials were significantly higher than the responses to reward omission in incorrect trials ($P < 0.05$, two-sample t test). Responses to the reward were measured in a 400-ms window centered 350 ms after the PB.

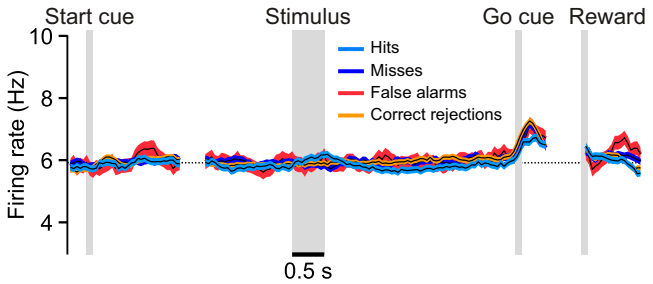

Fig. S2. Mean firing rate of the discarded neurons. Mean population firing rate (black line, ±SEM colored bands) of the discarded neurons was plotted as a function of time for the four trial types. Activity is aligned to the start cue (*Left*), the go cue (*Center*), and reward delivery (*Right*). The dotted line indicates the baseline activity (5.9 spikes per second). The color code used to indicate the four trial types is the same as in Fig. 1*B*.

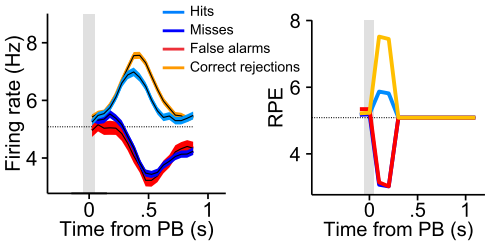

Fig. S3. DA phasic responses and RPEs at the reward delivery. Both the mean firing rate (*Left*) and the RPE (*Right*) showed a positive activation in rewarded trials and a pause in incorrect decision trials. The larger fraction of rewarded trials with the stimulus-present decision was responsible for the smaller RPE in hit trials than in CR ones (*Right*). The color code used to indicate the four trial types is the same as in Fig. 1*B*. PB denotes the push button event.

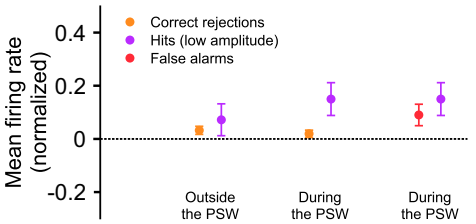

Fig. S4. DA activity in low-amplitude hit trials compared with the activity in stimulus-absent trials. The mean activity in low-amplitude hit trials (*SI Materials and Methods*) exhibited a significant positive modulation with respect to CR trials during the PSW ($P < 0.05$, two-sample one-tailed t test) but not outside it ($P = 0.26$, two-sample one-tailed t test). Notably the activity in low-amplitude hit trials and in FA trials during the PSW did not show any significant difference ($P = 0.21$, two-sample one-tailed t test).

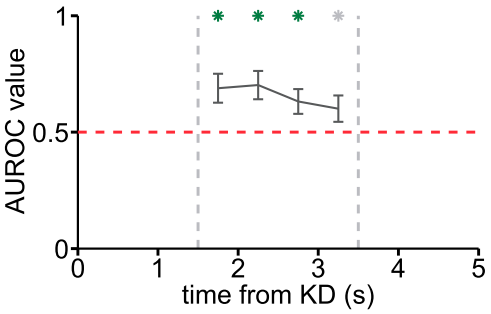

Fig. S5. The activity of DA neurons covaries with the animal's choice during the presentation of the stimulus. The PSW was divided into four temporal bins. Hit and miss trials of intermediate amplitudes were separately sorted according to their SO timing. For each time bin the normalized responses to the stimulus in hit and miss trials were used to evaluate AUROC values. The analysis showed that the DA activity covaried with behavior significantly during the first three time bins ($P < 0.01$). The small value of the index at the end of the PSW could be a consequence of the dynamics of cortical networks. Those dynamics can be explained (31) in terms of a response criterion that becomes smaller during the PWS (to improve detection). After this temporal window the criterion increases to reduce the production of FA events. It is reasonable to think that by the end of the PWS the criterion evolves continuously from a small to a large value. As a consequence during the last time bin the firing response of cortical neurons in miss trials is more similar to the response in hit trials; DA midbrain neurons reflect this situation. Green asterisks indicate significant AUROC values. The red dashed line indicates the chance level (AUROC = 0.5).

Table S1. List of the parameters adopted by the computational model

Component of the RL model	Description	Symbol	Value
Critic	Learning rate in acquisition	η_c^+	0.1
	Learning rate in extinction	η_c^-	0.2
	Rectification	ψ	0.15
	Discount factor	γ	0.98
	Smear of the $\boldsymbol{T}$ functions	k	80
	Spacing of the $\boldsymbol{T}$ functions	c	0.2
Actor	Learning rate in acquisition	η_a^+	0.03
	Learning rate in extinction	η_a^-	0.1
	Noise of the softmax	β	0.5
	Decay of stimulus trace	ρ	0.98