

Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex

Yuji K Takahashi¹, Matthew R Roesch^{2,3}, Robert C Wilson^{4,5}, Kathy Toreson¹, Patricio O'Donnell^{1,6}, Yael Niv^{4,5,8} & Geoffrey Schoenbaum^{1,6-8}

The orbitofrontal cortex has been hypothesized to carry information regarding the value of expected rewards. Such information is essential for associative learning, which relies on comparisons between expected and obtained reward for generating instructive error signals. These error signals are thought to be conveyed by dopamine neurons. To test whether orbitofrontal cortex contributes to these error signals, we recorded from dopamine neurons in orbitofrontal-lesioned rats performing a reward learning task. Lesions caused marked changes in dopaminergic error signaling. However, the effect of lesions was not consistent with a simple loss of information regarding expected value. Instead, without orbitofrontal input, dopaminergic error signals failed to reflect internal information about the impending response that distinguished externally similar states leading to differently valued future rewards. These results are consistent with current conceptualizations of orbitofrontal cortex as supporting model-based behavior and suggest an unexpected role for this information in dopaminergic error signaling.

Midbrain dopamine neurons signal errors in reward prediction¹⁻³. These error signals are required for learning in a variety of theoretical accounts⁴⁻⁶. By definition, calculation of these errors requires information about the value of the rewards expected in a given circumstance or 'state'. In temporal difference reinforcement learning (TDRL) models, such learned expectations contribute to computations of prediction errors and are modified on the basis of these errors. However the neural source of this expected value signal has not been established for dopamine neurons in the ventral tegmental area (VTA). Here we tested whether one contributor might be the orbitofrontal cortex (OFC), a prefrontal area previously shown to be critical for using information about the value of expected rewards to guide behavior⁷⁻¹¹.

RESULTS

To test whether OFC contributes to reward prediction errors, we recorded single-unit activity from putative dopamine neurons in the VTA in rats with ipsilateral sham ($n = 6$) or neurotoxic lesions ($n = 7$) of OFC (Fig. 1). Lesions targeted the ventral and lateral orbital and ventral and dorsal agranular insular areas in the bank of the rhinal sulcus, resulting in frank loss of neurons in 33.4% (23–40%) of this layered cortical region across the seven subjects (Fig. 1c, inset). Neurons in this region fire in anticipation of an expected reward¹² and interact with VTA to drive learning in response to prediction errors¹³. Notably, sparse direct projections from this part of OFC to VTA are largely unilateral¹⁴, and neither direct nor indirect input to VTA

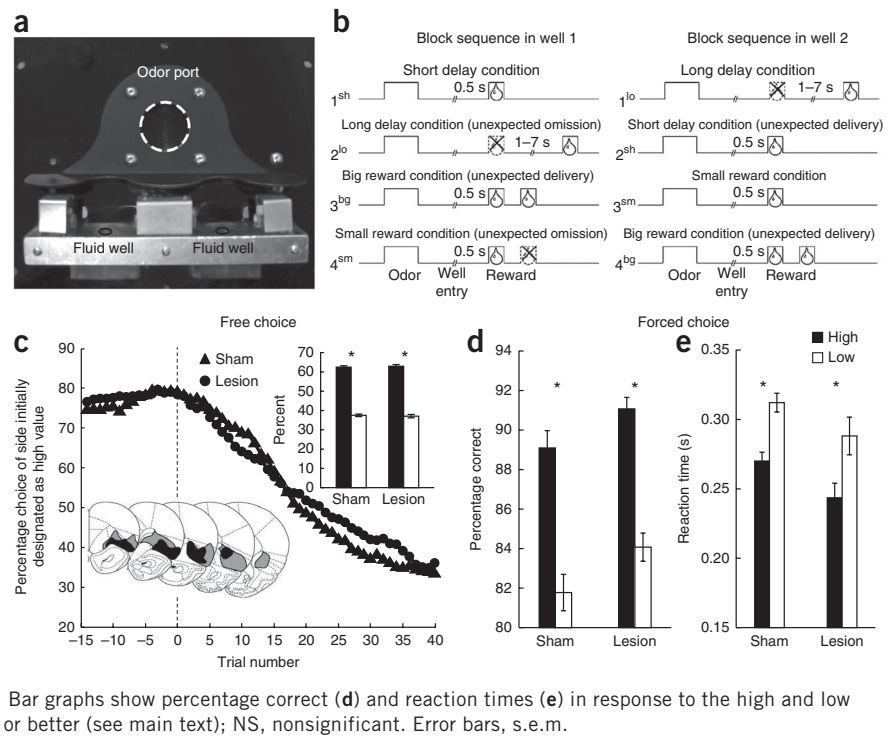
from contralateral OFC is sufficient to support normal learning¹³. Therefore, ipsilateral lesions should severely diminish any influence of OFC signaling on VTA in the lesioned hemisphere while leaving the circuit intact in the opposite hemisphere to avoid confounding behavioral deficits.

Neurons were recorded in an odor-guided choice task used previously to characterize signaling of errors and outcome expectancies^{12,13,15}. On each trial, rats responded at one of two adjacent wells after sampling one of three different odor cues at a central port (Fig. 1a). One odor signaled sucrose reward in the right well (forced choice right), a second odor signaled sucrose reward in the left well (forced choice left), and a third odor signaled the availability of reward at either well (free choice). To generate errors in the prediction of rewards, we manipulated the timing (Fig. 1b, blocks 1 and 2) or size of the reward (Fig. 1b, blocks 3 and 4) across blocks of trials. This resulted in the introduction of new and unexpected rewards, when immediate or large rewards were instituted at the start of blocks 2^{sh}, 3^{bg}, 4^{bg} (Fig. 1b), and omission of expected rewards, when delayed or small rewards were instituted at the start of blocks 2^{lo} and 4sm (Fig. 1b). Sh, lo, bg and sm indicate short delay, long delay, big reward and small reward conditions, respectively.

As expected, sham-lesioned rats changed their choice behavior across blocks in response to the changing rewards, choosing the higher value reward more often on free-choice trials (t -test, $t_{100} = 18.91$, $P < 0.01$; Fig. 1c, inset) and responding more accurately (t -test, $t_{100} = 10.77$, $P < 0.01$; Fig. 1d) and with shorter reaction times (t -test, $t_{100} = 13.32$,

¹Department of Anatomy and Neurobiology, University of Maryland School of Medicine, Baltimore, Maryland, USA. ²Department of Psychology, University of Maryland College Park, College Park, Maryland, USA. ³Program in Neuroscience and Cognitive Science, University of Maryland College Park, College Park, Maryland, USA. ⁴Department of Psychology, Princeton University, Princeton, New Jersey, USA. ⁵Neuroscience Institute, Princeton University, Princeton, New Jersey, USA. ⁶Department of Psychiatry, University of Maryland School of Medicine, Baltimore, Maryland, USA. ⁷National Institute on Drug Abuse Intramural Research Program, Baltimore, Maryland, USA. ⁸These authors contributed equally to this work. Correspondence should be addressed to Y.K.T. (ytaka001@umaryland.edu) or G.S. (schoenbg@schoenbaumlab.org).

Figure 1 Apparatus and behavioral results. (a) Odor port and fluid wells. (b) Time course of stimuli (odors and rewards) presented to the rat on each trial. At the start of each recording session (block 1), one well was randomly designated as short (a 0.5 s delay before reward) and the other, long (a 1–7 s delay before reward). In the second block of trials, these contingencies were switched (block 2). In blocks 3 and 4, the delay was held constant while the number of the rewards delivered was manipulated. Expected rewards were thus omitted on long and small trials at the start of blocks 2 (2^{lo}) and 4 (4^{sm}), respectively, and rewards were delivered unexpectedly on short and big trials at the start of blocks 2 (2^{sh}) and 3 and 4 (3^{bg} and 4^{bg}), respectively. (c) Choice behavior in trials before and after the switch from high-valued outcome (averaged across short and big) to a low-valued outcome (averaged across long and small). Inset bar graphs show average percentage choice for high-value (black) versus low-value (white) outcomes across all free-choice trials. Inset brain sections illustrate the extent of the maximum (gray) and minimum (black) lesion at each level in OFC in the lesioned rats. (d,e) Behavior on forced-choice trials. Bar graphs show percentage correct (d) and reaction times (e) in response to the high and low value cues across all recording sessions. * $P < 0.05$ or better (see main text); NS, nonsignificant. Error bars, s.e.m.



$P < 0.01$; **Fig. 1e**) on forced-choice trials when the high value reward was at stake. Rats with unilateral OFC lesions showed similar behavior (t -test: percent choice, $t_{84} = 14.51$, $P < 0.01$; percent correct, $t_{84} = 9.88$, $P < 0.01$; reaction time, $t_{84} = 8.32$, $P < 0.01$; **Fig. 1c–e**), and direct comparisons of all three performance measures across groups revealed no significant differences (ANOVA, sham versus lesioned; percent choice, $F_{1,184} = 0.16$, $P = 0.68$; percent correct, $F_{2,183} = 2.11$, $P = 0.12$; reaction time, $F_{2,183} = 2.92$, $P = 0.06$).

We identified dopamine neurons in the VTA by means of a cluster analysis based on spike duration and amplitude ratio (**Fig. 2**). Although the use of such criteria has been questioned¹⁶, the particular analysis used here isolates neurons whose firing is sensitive to intravenous infusion of apomorphine¹⁵ or quinpirole¹⁷. Additionally, neurons identified by this cluster analysis are selectively activated by optical stimulation in tyrosine hydroxylase–channelrhodopsin-2 mutants¹⁷ and show reduced bursting in tyrosine hydroxylase–striatal-specific NMDAR1 knockouts¹⁷. Although these criteria may exclude some dopamine neurons, only neurons in this cluster signaled reward prediction errors in appreciable numbers in our previous work¹⁵.

This analysis identified 52 of 481 recorded neurons as dopaminergic in shams (**Fig. 2a**) and 76 of 500 as dopaminergic in OFC-lesioned rats (**Fig. 2b**). These neurons had spike durations and amplitude ratios that differed significantly (>3 s.d.) from those of other neurons. Of these, 30 in sham and 50 in OFC-lesioned rats increased firing in response to reward (compared with baseline during the inter-trial interval; t -test, $P < 0.05$; proportions did not differ in sham versus lesioned: chi-squared test, $\chi^2 = 0.86$, degrees of freedom (d.f.) = 1, $P = 0.35$). There were no apparent effects of OFC lesions on the waveform characteristics of these neurons (**Fig. 2c**, t -test: amplitude ratio, $t_{127} = 0.53$, $P = 0.59$; duration, $t_{127} = 0.78$, $P = 0.43$). The average baseline activity of reward-responsive and nonresponsive dopamine neurons, taken during the 500 ms before the light onset that signaled start of a trial, was also similar in the two groups (**Fig. 2f**; sham versus lesioned, t -test: reward-responsive dopamine neurons, $t_{78} = 0.49$, $P = 0.62$; reward-nonresponsive dopamine neurons, $t_{46} = 1.57$, $P = 0.12$),

as was the distribution of the baseline firing (**Fig. 2d,e**; sham versus lesioned, Wilcoxon: reward-responsive dopamine neurons, $P = 0.86$; reward-nonresponsive dopamine neurons, $P = 0.09$). Thus, OFC lesions did not affect the firing of dopaminergic neurons (also see **Supplementary Fig. 1**). Of note, non-dopaminergic neurons fired significantly more slowly in the OFC-lesioned rats (sham versus lesioned, t -test: $t_{851} = 3.81$, $P < 0.01$; see **Supplementary Fig. 2**).

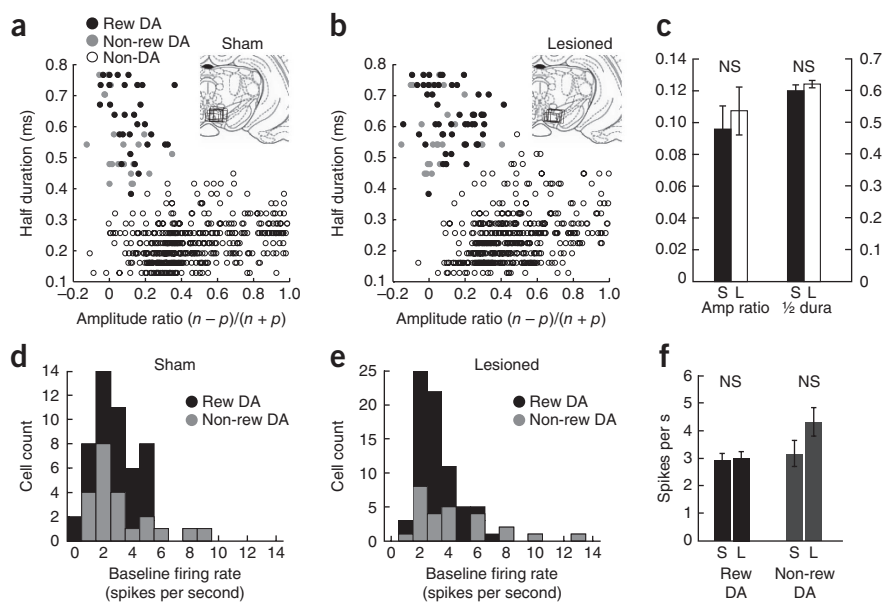
OFC supports dopaminergic error signals

Previous work has shown that prediction-error signaling is largely restricted to reward-responsive dopamine neurons¹⁵ (see **Supplementary Figs. 1 and 2** for analysis of other populations). As expected, activity in these neurons in sham-lesioned rats increased in response to unexpected reward and decreased in response to omission of an expected reward. As seen from unit examples and population responses (**Fig. 3a**), neural activity increased when a new reward was introduced (start of block 2^{sh}) and decreased when an expected reward was omitted (start of block 2^{lo}). In both cases, the change in activity was maximal at the beginning of the block and then diminished with learning.

These patterns were substantially muted in OFC-lesioned rats. Although dopamine neurons still showed phasic firing to unexpected rewards, this response was not as pronounced at the beginning of the block, nor did it change substantially with learning. In addition, the suppression of activity normally caused by unexpected reward omission was largely abolished (**Fig. 3b**).

These effects are quantified in **Figure 3c,d**, which plots the average activity across all reward-responsive dopamine neurons in each group, on each of the first and last ten trials in all blocks in which we delivered a new, unexpected reward (blocks 2^{sh} , 3^{bg} , 4^{bg}) or omitted an expected reward (blocks 2^{lo} and 4^{sm}). In sham-lesioned rats, reward-responsive dopamine neurons increased firing upon introduction of a new reward and suppressed firing on omission of an expected reward. In each case, the change in firing was maximal on the first trial and diminished significantly thereafter (**Fig. 3c**). Two-factor ANOVAs

Figure 2 Identification, waveform features and firing rates of putative dopamine and non-dopamine neurons. (a,b) Results of cluster analysis based on the half time of the spike duration and the ratio comparing the amplitude of the first positive and negative waveform segments ($(n-p)/(n+p)$). The center and variance of each cluster was computed without data from the neuron of interest, and then that neuron was assigned to a cluster if it was within 3 s.d. of the cluster's center. Neurons that met this criterion for more than one cluster were not classified. This process was repeated for each neuron. Reward-responsive dopamine neurons (rew DA), black; reward-nonresponsive dopamine neurons (non-rew DA), gray; neurons that classified with other clusters, no clusters or more than one cluster, open circles. Insets in each panel indicate location of the electrode tracks in sham (a) and OFC-lesioned rats (b). (c) Bar graphs indicating average amplitude (amp) ratio and half duration (1/2 dura) of putative dopamine neurons in sham (S) and OFC-lesioned rats (L). (d-f) Distribution and average baseline firing rates for reward-responsive (black) and nonresponsive (gray) dopamine neurons in sham (d,f) and OFC-lesioned rats (e,f). NS, nonsignificant (see main text). Error bars, s.e.m.



comparing firing to unexpected reward (or reward omission) to background firing (average firing during inter-trial intervals) showed significant interactions between trial period and trial number in each case (reward versus background, $F_{19,532} = 4.37$, $P < 0.0001$; omission versus background, $F_{19,532} = 3.57$, $P < 0.0001$). *Post hoc* comparisons showed that activity on the first five trials differed significantly from background, as well as from activity on later trials, for both unexpected reward and reward omission (P values < 0.01). Furthermore, the distribution of difference scores comparing each neuron's firing early and late in the block was shifted significantly above zero for unexpected reward (Fig. 3e; Wilcoxon signed-rank test, $P < 0.01$) and below zero for reward omission (Fig. 3f; Wilcoxon signed-rank test, $P < 0.01$), and there was a significant inverse correlation between changes in firing in response to unexpected reward and reward omission (Fig. 3g; $r = 0.43$, $P < 0.05$). These results are consistent with bidirectional prediction error signaling in the individual neurons at the time of reward in the sham-lesioned rats.

By contrast, the activity of reward-responsive dopamine neurons in OFC-lesioned rats did not change substantially across trials in response to reward omission (Fig. 3d). Two-factor ANOVAs comparing these data to background firing revealed a main effect of reward ($F_{1,48} = 46.3$, $P < 0.0001$) but no effect of omission nor any interactions with trial number (F values < 1.29 , P values > 0.17), and *post hoc* comparisons showed that the reward-evoked response was significantly higher than background on every trial in Figure 3d (P values < 0.01), whereas the omission-evoked response did not differ on any trial (P values > 0.05).

Examination of the difference scores across individual neurons in OFC-lesioned rats showed similar effects. For example, although the distribution of these scores was shifted significantly above zero for unexpected reward (Fig. 3h; Wilcoxon signed-rank test, $P < 0.01$), the shift was significantly less than that in shams (Fig. 3e versus Fig. 3h; Mann-Whitney U test, $P < 0.001$), as was the actual number of individual neurons in which reward-evoked activity declined significantly with learning (Fig. 3e versus Fig. 3h; chi-squared test, $\chi^2 = 5.12$, d.f. = 1, $P = 0.02$). Furthermore, not a single neuron in the lesioned rats suppressed firing significantly in response to reward omission (Fig. 3i), and the

distribution of these scores was less shifted than in shams (Fig. 3f versus Fig. 3i; Mann-Whitney U test, $P < 0.001$) and did not differ from zero (Fig. 3i; Wilcoxon signed-rank test, $P = 0.12$). There was no significant inverse correlation between changes in firing in response to unexpected reward and to reward omission (Fig. 3j; $r = 0.10$, $P = 0.47$).

Thus, ipsilateral lesions of OFC substantially diminished the normal effect of learning on firing in response to unexpected reward and reward omission in VTA dopamine neurons. This effect was observed even though the rats' behavior indicated that they learned to expect reward at the same rate as controls (see Fig. 1). These results, along with a parallel analysis of activity at the time of delivery of the delayed reward in blocks 1 and 2 (see Supplementary Fig. 3), all point to a critical contribution of OFC to the prediction errors signaled by VTA dopamine neurons at the time of reward.

According to prevailing frameworks such as TDRL, prediction error signals should also be evident in response to cues. Consistent with this, reward-responsive dopamine neurons in sham-lesioned rats responded phasically during and immediately after sampling of the odors, and this phasic response differed according to the expected value of the trial (Fig. 4). Thus, on forced-choice trials, the average firing rate was higher during (and immediately after) sampling of the high value cue than during sampling of the low value cue (Fig. 4a). This difference was not present in the initial trials of a block but rather developed with learning. A two-factor ANOVA comparing firing to the odor cues across all neurons showed a significant main effect of value ($F_{1,28} = 12.2$, $P < 0.01$) and a significant interaction between value and learning ($F_{1,28} = 18.0$, $P < 0.001$). We also quantified the effect of value by calculating the difference in firing to the high and low value cues for each neuron before and after learning; the distribution of this score was shifted significantly above zero after (Fig. 4a, late distribution; Wilcoxon signed-rank test, $P < 0.01$) but not before learning (Fig. 4a, early distribution; Wilcoxon signed-rank test, $P = 0.68$).

This pattern was also evident on free-choice trials, in which a single odor cue was presented but either of the two rewards could be selected by responding to the appropriate well. Dopaminergic activity in sham-lesioned rats increased during sampling of the single cue and

then diverged in accordance with the future choice of the rat, increasing more before selection of the high value well than the low value well. ANOVA comparing firing between odor offset (when the rat was still in the odor port) and earliest well response confirmed this effect (Fig. 4c; $F_{1,28} = 8.33$, $P < 0.01$), as did the distribution of the difference scores comparing firing during this period on high minus low value trials for each neuron (Fig. 4c; Wilcoxon signed-rank test, $P < 0.01$).

These cue-evoked effects were also altered in OFC-lesioned rats. On forced-choice trials, reward-responsive dopamine neurons fired differentially based on cue value (Fig. 4b; two-factor ANOVA: significant

main effect of value, $F_{1,48} = 10.4$, $P < 0.01$; significant interaction between value and learning, $F_{1,48} = 6.36$, $P < 0.05$), and the distribution of the difference scores comparing firing to the high and low value cues after learning was shifted significantly above zero (Fig. 4b, late distribution; Wilcoxon signed-rank test, $P < 0.01$). However, the differential firing in OFC-lesioned rats on forced-choice trials was weaker than in shams, and the number of neurons in which firing showed a significant effect of cue value was significantly lower in OFC-lesioned than sham-lesioned rats (Fig. 4a versus Fig. 4b; chi-squared test, $\chi^2 = 5.19$, d.f. = 1, $P = 0.02$).

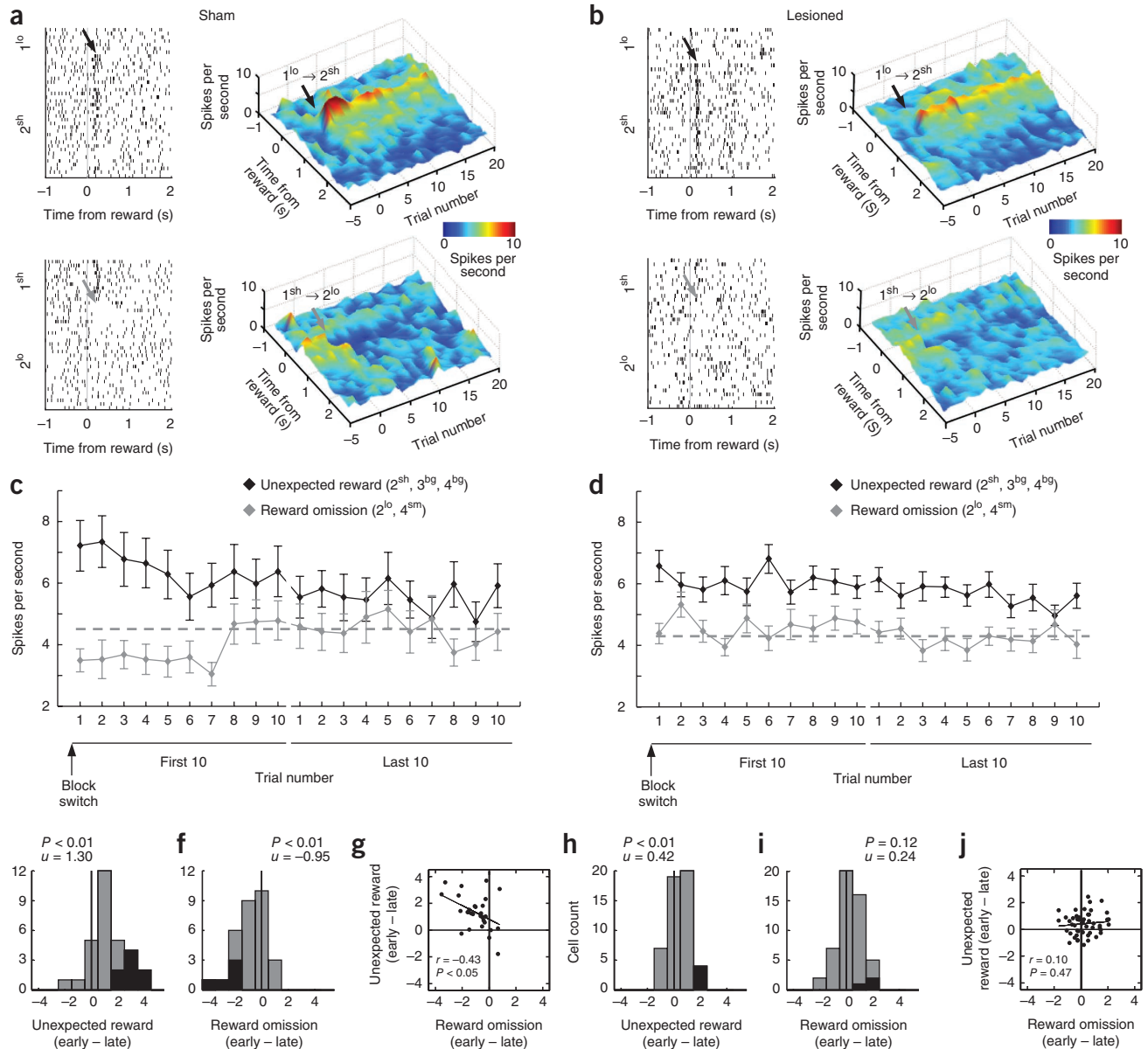
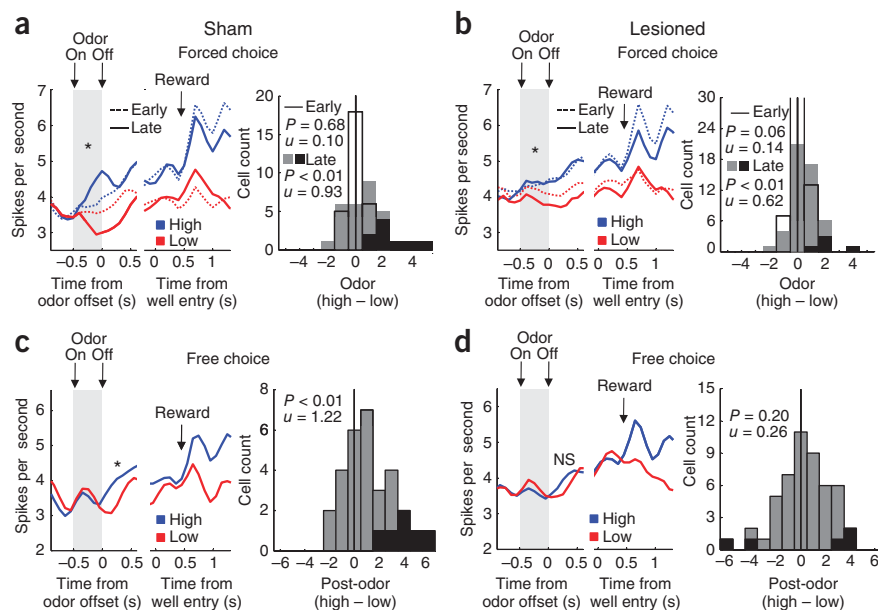


Figure 3 Changes in activity of reward-responsive dopamine neurons in response to unexpected reward delivery and omission. (a,b) Activity in a representative neuron (raster) or averaged across all reward-responsive dopamine neurons (heat plot) in sham (a) and OFC-lesioned rats (b) in response to introduction of unexpected reward in block 2^{sh} (top plots, black arrows) and omission of expected reward in block 2^{lo} (bottom plots, gray arrows). (c,d) Average firing during the period 500 ms after reward delivery or omission in reward-responsive dopamine neurons in sham (c) and OFC-lesioned rats (d) in blocks in which an unexpected reward was instituted (blocks 2^{sh}, 3^{bg} and 4^{bg}, black lines) or an expected reward omitted (blocks 2^{lo} and 4sm, gray lines). Dashed lines indicate background firing. Error bars, s.e.m. (e–j) Distribution of difference scores and scatter plots comparing firing to unexpected reward and reward omission early versus late in relevant trial blocks in sham (e–g) and OFC-lesioned rats (h–j). Difference scores were computed from the average firing rate of each neuron in the first 5 minus the last 15 trials in relevant trial blocks. Black bars represent neurons in which the difference in firing was statistically significant (*t*-test; $P < 0.05$). The numbers in upper left of each panel indicate results of Wilcoxon signed-rank test (P) and the average difference score (u).

Figure 4 Changes in activity of reward-responsive dopamine neurons during and after odor cue sampling on forced- and free-choice trials. (**a–d**) Neural activity during forced-choice (**a,b**) and free-choice (**c,d**) trials in shams (**a,c**) and OFC-lesioned rats (**b,d**). Line plots show average activity synchronized to odor offset or well entry across all blocks on trials involving the high and low value cues. * $P < 0.05$ or better on *post hoc* testing (see main text); NS, nonsignificant. The numbers indicate results of Wilcoxon signed-rank test (P) and the average difference score (u). In **a,b**, activity is shown separately for the first 15 (early) and the last 5 trials (late) in each block, corresponding to the time during and after learning in response to a change in the size or timing of reward. Histograms show the distribution of difference scores comparing firing during sampling of the high minus the low value cues, early (outlined bars) and late (filled bars) in the blocks. Black bars represent neurons in which the difference in firing in late trials was statistically significant (t -test; $P < 0.05$). In **c,d** histograms show the distribution of difference scores comparing firing between odor offset and earliest well response on high minus low value trials. Black bars represent neurons in which the difference in firing was statistically significant (t -test; $P < 0.05$).



In addition, on free-choice trials, the difference in firing that emerged after cue-sampling in sham-lesioned rats was wholly absent in OFC-lesioned rats (**Fig. 4d**); a two-factor ANOVA comparing firing during this post-cue-sampling period in OFC-lesioned rats with that in shams showed a significant interaction between group and value (**Fig. 4c** versus **Fig. 4d**; $F_{1,78} = 4.05$, $P < 0.05$), and *post hoc* testing showed that the significant difference present in shams was not present in lesioned rats (**Fig. 4d**; $F_{1,48} = 1.71$, $P = 0.2$).

OFC modulates dopaminergic activity *in vivo*

The data above suggest that OFC modulates the firing of VTA dopamine neurons. To test this directly, we recorded juxtacellularly from VTA neurons in anesthetized rats. We identified 15 neurons with amplitude ratios and spike durations similar to those of the putative dopamine neurons recorded in the awake, behaving rats. These neurons showed low baseline firing rates (3.54 ± 1.35 Hz) as well as bursting patterns characteristic of dopaminergic neurons¹⁸. Six stained with Neurobiotin and colocalized tyrosine hydroxylase (**Fig. 5a**).

Eleven (73.3%) showed a statistically significant suppression of firing during and immediately after electrical stimulation of the OFC (five-pulse, 20-Hz trains; **Fig. 5b,c**), including all six tyrosine hydroxylase-positive neurons (**Fig. 5c**, bottom). In each case, inhibition began during the 200-ms period of OFC stimulation and lasted for several hundred milliseconds, averaging 393.3 ± 184.9 ms (range 220–740 ms). Inhibition was sometimes followed by a rebound excitation. Inhibition was not observed during the inter-stimulation interval in these neurons and thus was a specific effect of OFC stimulation.

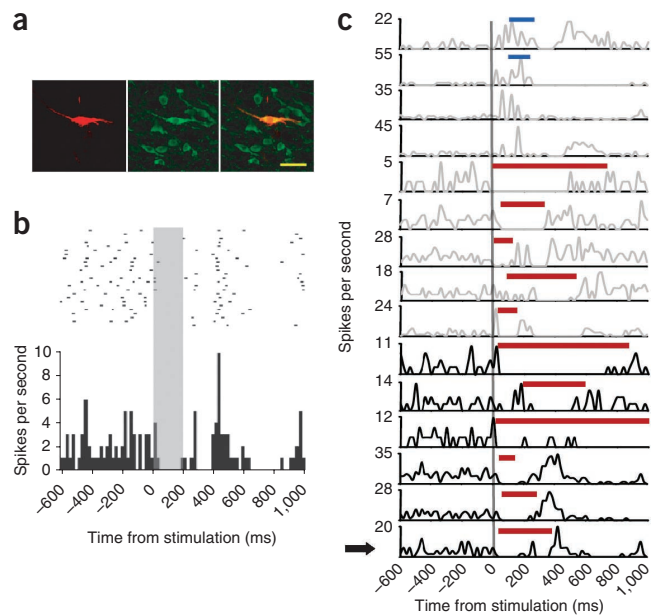
Of the four neurons that did not show a significant suppression, two showed a significant increase in firing in response to OFC stimulation, suggesting that OFC can excite as well as inhibit firing in dopamine neurons, whereas two showed significant suppression epochs only after the end of stimulation; these were considered nonresponsive. The average latency of onset of the OFC-dependent responses was 93.9 ± 106.9 ms (range, 0–980 ms).

OFC does not convey value to VTA

Our results show that the OFC contributes to intact error signaling by dopamine neurons in VTA. To understand the nature of this contribution, we used computational modeling. In all models, we used the TDRL framework⁴ that has been used extensively to describe animal learning in reward-driven tasks and the generation of phasic dopaminergic firing patterns¹⁹. In this framework a prediction error signal δ_t at time t is computed as $\delta_t = r_t + V(S_t) - V(S_{t-1})$, where r_t is the currently available reward (if any), S_t is the current state of the task (the available stimuli and so forth), $V(S_t)$ is the value of the current state—that is, the expected amount of future rewards—and $V(S_{t-1})$ is the value of the previous state; that is, the total predicted rewards before this time point. The prediction error is used to learn the state values through experience with the task, by increasing $V(S_{t-1})$ if the prediction error δ_t is positive (indicating that obtained and future expected rewards exceed the initial expectation) and decreasing $V(S_{t-1})$ if the prediction error is negative (indicating over-optimistic initial expectations that must be reduced). These prediction errors are the signals thought to be reported by dopamine neurons¹⁹ and were indeed well-matched to the neural data from VTA dopamine neurons recorded in sham-lesioned rats (**Fig. 6a**).

Owing to the involvement of OFC in signaling reward expectancies^{7–10}, we initially hypothesized that OFC might convey to dopamine neurons the value of states in terms of the expected future reward $V(S_t)$ at each point in time. However, modeling the OFC lesion by removing expected values from the calculation of prediction errors failed to replicate the experimental results (**Fig. 6b**, model 1; for details of this and subsequent models, see Online Methods). Specifically, although removal of (learned) values accurately predicted that firing in dopamine neurons at the time of unexpected reward or reward omission would remain unchanged with learning, this model could not account for the reduced initial response to unexpected rewards in OFC- versus sham-lesioned rats (**Fig. 3c** versus **Fig. 3d**), nor could it generate differential firing to the odor cues on forced-choice trials (**Fig. 4b**). Thus, a complete loss of value input to dopamine neurons did not reproduce the effects of OFC lesions on error signaling.

Figure 5 Changes in dopamine neuron activity in response to OFC stimulation. **(a)** An example of putative dopamine neuron labeled with Neurobiotin (left, red) and tyrosine hydroxylase (TH; middle, green). This neuron had morphological characteristics of dopamine neurons (bipolar dendritic orientation) and showed colocalization of Neurobiotin and TH (right, merged). Scale bar, 50 μm . **(b)** Raster plot and peri-stimulus histogram showing activity in the TH⁺ neuron from **a** before, during and after OFC stimulation. Top: each line represents a trial and dots indicate time of action potential firing; gray box indicates period of OFC stimulation. Bottom: cumulative histogram depicting firing across all trials and revealing a pause during the stimulation. **(c)** Firing rate plots showing activity in each of the 15 recorded neurons before, during and after OFC stimulation. Arrow indicates neuron shown in **a, b**. Each line shows the average firing rate per stimulation trial for a given neuron. Activity is aligned to onset of OFC stimulation (vertical gray line). Thirteen neurons showed periods of significant ($P < 0.001$; see Online Methods) inhibition (red bars) or excitation (blue bars) that began during the stimulation. Excluding secondary or rebound excitation or inhibition evident in the figure, these neurons did not show significant epochs elsewhere in the inter-stimulus interval (not shown). Gray rate plots, putative dopamine neurons; black rate plots, TH immunopositive neurons.



We next considered whether a partial loss of the value signal might explain the observed effects of OFC lesions. A partial loss might occur if the lesions were incomplete (as they were) or if another brain region, such as contralateral OFC and/or ventral striatum, were also providing value signals. Although this produced slower learning, it did not prevent asymptotically correct values from being learned for the low and high reward port, as well as for the two choices in the free-choice trials. This occurs because the remaining value-learning structures still update their estimates of values on the basis of ongoing prediction errors and are thus able to compensate for the loss of some of the value-learning neurons. Thus, according to this model, prediction errors to reward should still decline with training, and prediction errors to cues and choices should still increase with training, predictions at odds with the empirical data (**Fig. 6b**, model 2).

A second way a partial loss of the value signal might occur is if only some of the rats had lesions sufficient to prevent value learning, whereas others had enough intact neural tissue to support value learning. This would amount to a partial loss of values between, rather than

within, subjects (**Fig. 6b**, model 3). This model did accurately predict some features of the population data, such as diminished (but still significant) differential firing to the odor cues on forced-choice trials (**Fig. 4b**). However, it too failed to explain the reduced initial response to unexpected rewards in OFC- versus sham-lesioned rats (**Fig. 3c** versus **Fig. 3d**), and it could not explain the absence of differential firing as a result of future expectation of low or high rewards on free-choice trials (**Fig. 4d**). Moreover, this between-subjects account was at odds with the observation that none of the individual neurons in OFC-lesioned rats showed intact error signaling (**Fig. 3h–j**).

OFC signals state information to VTA

Our models thus did not support the hypothesis that the OFC conveys some or all information about expected reward value to VTA neurons. Models in which this value signal was completely or even partially

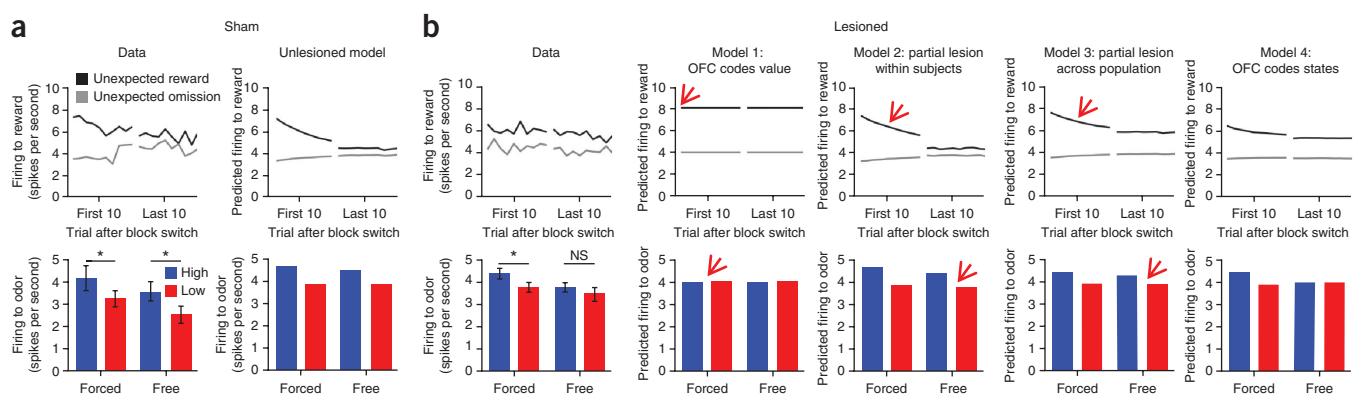


Figure 6 Comparison of model simulations and experimental data. **(a)** The unlesioned TDRL model and experimental data (\pm s.e.m.) from the sham-lesioned rats. Top: at the time of unexpected reward delivery or omission, the model predicts positive (black) and negative (gray) prediction errors whose magnitude diminishes as trials proceed. Bottom: at the time of the odor cue, the model reproduces the increased responding to high value (blue) relative to low value (red) cues on forced trials. Likewise, the model predicts differential firing at the time of decision on free-choice trials. **(b)** Lesioning the hypothesized OFC component in each model produces qualitatively different effects. Red arrows highlight discrepancies between the models and the experimental data, where these exist. Model 1, which postulates that OFC conveys expected values to dopamine neurons, cannot explain the reduced firing to unexpected rewards at the beginning of a block, nor can it reproduce the differential response to the two cues on forced-choice trials. Models 2 and 3, which assume a partial lesion of value encoding, cannot account for the lack of significant difference between high and low value choices on free-choice trials in the recorded data, and they incorrectly predict diminished responses at the time of reward after learning. Only model 4, in which OFC encoding enriches the state representation of the task by distinguishing between states on the basis of impending actions, was able to fully account for the results at the time of unexpected reward delivery or omission and at the time of odor presentation on free- and forced-choice trials.

removed had particular difficulty accounting for both the residual differential firing based on the learned value of the odor cues on forced-choice trials (Fig. 4b) and the loss of differential firing based on the value of the impending reward on free-choice trials (Fig. 4d). In each model, these were either both present or both absent.

The fundamental difference between forced-choice and free-choice trials is that in the former, two different cues were available to signal the two different expected values, whereas in the latter, signaling of the different values depended entirely on internal information regarding the rats' impending decision to respond in one direction or the other. Based on this distinction, we hypothesized that OFC might provide not state values *per se*, but rather more complex information about internal choices and their likely outcomes necessary to define precisely what state the task is in, particularly for states that are otherwise externally ambiguous (as is the case on free-choice trials). The provision of this information would allow other brain areas to derive more accurate value expectations for such states and to subsequently signal this information to VTA. Thus, in our fourth model, we hypothesized that the OFC provides input regarding state identity to the module that computes and learns the values, rather than directly to dopamine neurons. Removing the OFC would therefore leave an intact value learning system, albeit one forced to operate without some essential state information.

Consistent with this hypothesis, removing choice-based disambiguating input and leaving value learning to operate with more rudimentary, stimulus-bound states (Supplementary Fig. 4a) produced effects that closely matched empirical results from OFC-lesioned rats (Fig. 6b, model 4). Specifically, this model reproduced the patterns of cue selectivity evident in dopamine neurons in lesioned rats: on forced-choice trials, learning using only stimulus-bound states resulted in weaker differential prediction errors to the two odor cues; however, on free-choice trials, the lesioned model did not show divergent signaling at the decision point because it lacked the ability to use internal information about the impending choice to distinguish between the two decision states. Values for the two choices in the free-choice trials could not be learned no matter the size of the learning rate parameter or the duration of learning.

Notably, however, the consistency of this model with the neural data went beyond the effects on free-choice trials that motivated the model. In particular, the lesioned model showed firing to unexpected rewards and to reward omission that changed only very mildly through learning, which is similar to the neural data. Additionally, firing to an unexpected reward early in a block was lower than in the unlesioned model, again closely matching the neural results (Fig. 6b, data). Overall, this fourth model best captured the contribution of OFC to learning and prediction error signaling in our task (please see Supplementary Results for more discussion of modeling results).

DISCUSSION

Here we have shown that OFC is necessary for normal error signaling by VTA dopamine neurons. Dopamine neurons recorded in OFC-lesioned rats showed a muted increase in firing to an unexpected reward, and this firing failed to decline with learning as in sham-lesioned rats. These same neurons also failed to suppress firing when an expected reward was omitted, showed weaker differential firing to differently valued cues, and failed to show differential firing based on future expected rewards on free-choice trials. Computational modeling showed that while several of these features could be approximated by postulating that OFC provides predictive value information to support the computation of reward prediction errors, they were much better explained by an alternative model in which OFC was responsible for conveying information about impending actions to disambiguate

externally similar states leading to different outcomes. This suggests that rather than signaling expected values *per se*, the OFC might signal state information, thereby facilitating the derivation of more accurate values, particularly for states that are primarily distinguishable on the basis of internal rather than external information.

These results have important implications for understanding OFC and the role of VTA dopamine neurons in learning. Regarding OFC, these results provide a mechanism whereby information relevant to predicting outcomes, signaled by OFC neurons and critical to outcome-guided behavior, might influence learning²⁰. Although the involvement of OFC as a critical source of state representations is different from the role previously ascribed to OFC in learning (that of directly signaling expected reward values²⁰ or even prediction errors²¹), it would explain more clearly why this area is important for learning in some situations but not others, inasmuch as the situations requiring OFC, such as over-expectation and rapid reversal learning^{8,13,22,23}, are ones likely to benefit from disambiguation of similar states that lead to different outcomes. In these behavioral settings, optimal performance would be facilitated by the ability to create new states based on internally available information (that is, recognition that contingencies have changed)²⁴. Recent models suggest that state representations of tasks are themselves learned^{25,26}. Whether OFC is necessary for this learning process is not clear, but our results show that OFC is key for representing the resulting states. This idea is consistent with findings that OFC neurons encode all aspects of a task in a distributed and complex manner^{27–29} and with data showing that the OFC is particularly important for accurately attributing rewards to preceding actions^{30,31}, as this depends critically on representation of previous choices. In this regard, it is worth noting that OFC neurons have been shown to signal outcome expectancies in a response-dependent fashion in this and other behavioral settings^{12,32–36}.

The proposed contribution of OFC is also complementary to proposals that other brain regions, especially the ventral striatum, are important for value learning in TDRL models³⁷. OFC has strong projections to ventral striatum³⁸. Thus, information from OFC may facilitate accurate value signals in the ventral striatum, which might then be transmitted to midbrain dopaminergic neurons through inhibitory projections to contribute to prediction error signaling. Such a relay would seem essential to explain how glutamatergic output from the OFC acts to inhibit activity in VTA dopamine neurons, as demonstrated here and elsewhere³⁹. Other potential relays might include rostromedial tegmental nucleus, lateral habenula or even GABAergic interneurons in VTA, all of which receive input from OFC and can act to inhibit VTA dopamine neurons. Notably non-dopaminergic neurons in VTA, many of which are likely to be GABAergic, did show significantly lower baseline firing rates in OFC-lesioned rats than in controls (sham versus lesioned, *t*-test; $t_{851} = 3.81$, $P < 0.01$; Supplementary Fig. 2). These different pathways are not mutually exclusive, and each would be consistent with the long-latency, primarily inhibitory effects of OFC stimulation on dopamine activity shown here *in vivo*.

Finally, these results expand the potential role of VTA dopamine neurons in learning, by showing that the teaching signals encoded by these neurons are based, in part, on prefrontal representations. These prefrontal representations are critical for goal-directed or model-based behaviors⁴⁰; OFC in particular is necessary for changes in conditioned responses after reinforcer devaluation and other behaviors^{7,8,41,42} that require knowledge of how different states (cues, responses and decisions, rewards, and internal consequences) are linked together in a task. However, with the exception of two recent reports^{43,44}, this knowledge has not been thought to contribute to the so-called cached values underlying dopaminergic errors. Our results

show that these prefrontal representations do contribute to the value signal used by dopamine neurons to calculate errors. Correlates with action sequences, inferred values and impending actions evident in recent dopamine recording studies could derive from access to these states and the transitions between them thought to reside in orbital and other prefrontal areas^{45,46}. Full access to model-based task representations—the states, transition functions and derived values—would expand the types of learning that might involve dopaminergic error signals to complex associative settings^{47–49} more likely to reflect situations in the real world.

METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/natureneuroscience/>.

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS

This work was supported by grants from the US National Institute on Drug Abuse to G.S. and M.R. and from the US National Institute on Mental Health to Y.K.T., a Sloan Research Fellowship to Y.N. and a Binational United States-Israel Science Foundation grant to Y.N. and R.C.W.

AUTHOR CONTRIBUTIONS

G.S., M.R.R. and Y.K.T. conceived the initial unit recording study in awake rats; Y.K.T. and M.R.R. carried it out, and Y.K.T. and G.S. analyzed the data. Subsequently, G.S. approached P.O. and Y.N. regarding *in vivo* recording and computational modeling, respectively. K.T. conducted the *in vivo* experiments, and K.T. and P.O. analyzed the data. R.C.W. and Y.N. conceived the alternative computational models, R.C.W. carried out the modeling, and Y.N. and R.C.W. interpreted the experimental data in light of simulation results. Y.N., G.S. and Y.K.T. collaborated in writing the manuscript with assistance from the other team members.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/natureneuroscience/>.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Hollerman, J.R. & Schultz, W. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* **1**, 304–309 (1998).
- Pan, W.-X., Schmidt, R., Wickens, J.R. & Hyland, B.I. Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. *J. Neurosci.* **25**, 6235–6242 (2005).
- Bayer, H.M. & Glimcher, P. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141 (2005).
- Sutton, R.S. & Barto, A.G. *Reinforcement Learning: An Introduction*. (MIT Press, 1998).
- Pearce, J.M. & Hall, G. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* **87**, 532–552 (1980).
- Rescorla, R.A. & Wagner, A.R. in *Classical Conditioning II: Current Research and Theory* (eds. Black, A.H. & Prokasy, W.F.) 64–99 (Appleton-Century-Crofts, 1972).
- Pickens, C.L. *et al.* Different roles for orbitofrontal cortex and basolateral amygdala in a reinforcer devaluation task. *J. Neurosci.* **23**, 11078–11084 (2003).
- Izquierdo, A., Suda, R.K. & Murray, E.A. Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. *J. Neurosci.* **24**, 7540–7548 (2004).
- O'Doherty, J.P., Deichmann, R., Critchley, H.D. & Dolan, R.J. Neural responses during anticipation of a primary taste reward. *Neuron* **33**, 815–826 (2002).
- Gottfried, J.A., O'Doherty, J. & Dolan, R.J. Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* **301**, 1104–1107 (2003).
- Padoa-Schioppa, C. & Assad, J.A. Neurons in orbitofrontal cortex encode economic value. *Nature* **441**, 223–226 (2006).
- Roesch, M.R., Taylor, A.R. & Schoenbaum, G. Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation. *Neuron* **51**, 509–520 (2006).
- Takahashi, Y.K. *et al.* The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. *Neuron* **62**, 269–280 (2009).
- Vázquez-Borsetti, P., Cortes, R. & Artigas, F. Pyramidal neurons in rat prefrontal cortex projecting to ventral tegmental area and dorsal raphe nucleus express 5-HT_{2A} receptors. *Cereb. Cortex* **19**, 1678–1686 (2009).
- Roesch, M.R., Calu, D.J. & Schoenbaum, G. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat. Neurosci.* **10**, 1615–1624 (2007).
- Margolis, E.B., Lock, H., Hjelmstad, G.O. & Fields, H.L. The ventral tegmental area revisited: is there an electrophysiological marker for dopaminergic neurons? *J. Physiol. (Lond.)* **577**, 907–924 (2006).
- Jin, X. & Costa, R.M. Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature* **466**, 457–462 (2010).
- Grace, A.A. & Bunney, B.S. The control of firing pattern in nigral dopamine neurons: burst firing. *J. Neurosci.* **4**, 2877–2890 (1984).
- Schultz, W., Dayan, P. & Montague, P.R. A neural substrate for prediction and reward. *Science* **275**, 1593–1599 (1997).
- Schoenbaum, G., Roesch, M.R., Stalnaker, T.A. & Takahashi, Y.K. A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nat. Rev. Neurosci.* **10**, 885–892 (2009).
- Sul, J.H., Kim, H., Huh, N., Lee, D. & Jung, M.W. Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. *Neuron* **66**, 449–460 (2010).
- Chudasama, Y. & Robbins, T.W. Dissociable contributions of the orbitofrontal and infralimbic cortex to pavlovian autoshaping and discrimination reversal learning: further evidence for the functional heterogeneity of the rodent frontal cortex. *J. Neurosci.* **23**, 8771–8780 (2003).
- Fellows, L.K. & Farah, M.J. Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. *Brain* **126**, 1830–1837 (2003).
- Gershman, S.J. & Niv, Y. Learning latent structure: carving nature at its joints. *Curr. Opin. Neurobiol.* **20**, 251–256 (2010).
- Redish, A.D., Jensen, S., Johnson, A. & Kurth-Nelson, Z. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychol. Rev.* **114**, 784–805 (2007).
- Gershman, S.J., Blei, D.M. & Niv, Y. Time, context and extinction. *Psychol. Rev.* **117**, 197–209 (2010).
- Ramus, S.J. & Eichenbaum, H. Neural correlates of olfactory recognition memory in the rat orbitofrontal cortex. *J. Neurosci.* **20**, 8199–8208 (2000).
- van Duuren, E., Lankelma, J. & Pennartz, C.M.A. Population coding of reward magnitude in the orbitofrontal cortex of the rat. *J. Neurosci.* **28**, 8590–8603 (2008).
- van Duuren, E. *et al.* Single-cell and population coding of expected reward probability in the orbitofrontal cortex of the rat. *J. Neurosci.* **29**, 8965–8976 (2009).
- Walton, M.E., Behrens, T.E.J., Buckley, M.J., Rudebeck, P.H. & Rushworth, M.F.S. Separable learning systems in the macaque brain and the role of the orbitofrontal cortex in contingent learning. *Neuron* **65**, 927–939 (2010).
- Tsuchida, A., Doll, B.B. & Fellows, L.K. Beyond reversal: a critical role for human orbitofrontal cortex in flexible learning from probabilistic feedback. *J. Neurosci.* **30**, 16868–16875 (2010).
- Tsujimoto, S., Genovesio, A. & Wise, S.P. Monkey orbitofrontal cortex encodes response choices near feedback time. *J. Neurosci.* **29**, 2569–2574 (2009).
- Feierstein, C.E., Quirk, M.C., Uchida, N., Sosulski, D.L. & Mainen, Z.F. Representation of spatial goals in rat orbitofrontal cortex. *Neuron* **51**, 495–507 (2006).
- Furuyashiki, T., Holland, P.C. & Gallagher, M. Rat orbitofrontal cortex separately encodes response and outcome information during performance of goal-directed behavior. *J. Neurosci.* **28**, 5127–5138 (2008).
- Abe, H. & Lee, D. Distributed coding of actual and hypothetical outcomes in the orbital and dorsolateral prefrontal cortex. *Neuron* **70**, 731–741 (2011).
- Young, J.J. & Shapiro, M.L. Dynamic coding of goal-directed paths by orbital prefrontal cortex. *J. Neurosci.* **31**, 5989–6000 (2011).
- O'Doherty, J. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004).
- Voorn, P., Vanderschuren, L.J.M.J., Groenewegen, H.J., Robbins, T.W. & Pennartz, C.M.A. Putting a spin on the dorsal-ventral divide of the striatum. *Trends Neurosci.* **27**, 468–474 (2004).
- Lodge, D.J. The medial prefrontal and orbitofrontal cortices differentially regulate dopamine system function. *Neuropsychopharmacology* **36**, 1227–1236 (2011).
- Daw, N.D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
- Burke, K.A., Franz, T.M., Miller, D.N. & Schoenbaum, G. The role of the orbitofrontal cortex in the pursuit of happiness and more specific rewards. *Nature* **454**, 340–344 (2008).
- Ostlund, S.B. & Balleine, B.W. Orbitofrontal cortex mediates outcome encoding in Pavlovian but not instrumental learning. *J. Neurosci.* **27**, 4819–4825 (2007).
- Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P. & Dolan, R.J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* (in the press).
- Simon, D.A. & Daw, N.D. Neural correlates of forward planning in a spatial decision task in humans. *J. Neurosci.* (in the press).
- Bromberg-Martin, E.S., Matsumoto, M., Hong, S. & Hikosaka, O. A pallidus-habenula-dopamine pathway signals inferred stimulus values. *J. Neurophysiol.* **104**, 1068–1076 (2010).
- Morris, G., Nevet, A., Arkadir, D., Vaadia, E. & Bergman, H. Midbrain dopamine neurons encode decisions for future action. *Nat. Neurosci.* **9**, 1057–1063 (2006).
- Hampton, A.N., Bossaerts, P. & O'Doherty, J.P. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* **26**, 8360–8367 (2006).
- Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J.P. Prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
- McDannald, M.A., Lucantonio, F., Burke, K.A., Niv, Y. & Schoenbaum, G. Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *J. Neurosci.* **31**, 2700–2705 (2011).

ONLINE METHODS

Behavioral and single-unit recording methods. *Subjects.* Thirteen male Long-Evans rats (Charles Rivers, ages 4–6 months) were tested at the University of Maryland School of Medicine in accordance with the University of Maryland School of Medicine Animal Care and Use Committee and US National Institutes of Health guidelines.

Surgical procedures. Recording electrodes were surgically implanted under stereotaxic guidance in the one hemisphere of VTA (5.2 mm posterior to bregma, 0.7 mm lateral and 7.0 mm ventral, angled 5° toward the midline from vertical). Some rats ($n = 7$) also received neurotoxic lesions of ipsilateral OFC by infusing NMDA (12.5 mg ml⁻¹) at four sites in each hemisphere: at 4.0 mm anterior to bregma, 3.8 mm ventral to the skull surface, 2.2 mm (0.1 μl) and 3.7 mm (0.1 μl) lateral to the midline; and at 3.0 mm anterior to bregma, 5.2 mm ventral to the skull surface, 3.2 mm (0.05 μl) and 4.2 mm lateral to the midline (0.1 μl). Controls ($n = 6$) received sham lesions in which burr holes were drilled and the pipette tip lowered into the brain but no solution delivered.

Behavioral task, single-unit recording, statistical analyses. Unit recording and behavioral procedures were identical to those described previously¹⁵. Statistical analyses are described in the main text.

Juxtacellular recording methods. *Subjects.* Nine male Long-Evans rats (Charles Rivers) were tested on postnatal day 60 at the University of Maryland School of Medicine in accordance with the University and US National Institutes of Health guidelines.

Surgical and recording procedures. Rats were anesthetized with chloral hydrate (400 mg per kilogram body weight, intraperitoneal) and placed in a stereotaxic apparatus. A bipolar concentric stimulating electrode was placed in the OFC (3.2 mm anterior and 3.0 mm lateral to bregma, and 5.2 mm ventral to the brain surface) connected to an Isoflex stimulus isolation unit and driven by a Master-8 stimulator (A.M.P.I.). Electrical stimulation of the OFC consisted of a five-pulse, 20-Hz train delivered every 10 s (pulse duration 0.5 ms, pulse amplitude 500 μA). Recording electrodes (resistance 10–25 MΩ) were filled with a 0.5 M NaCl, 2% (vol/vol) Neurobiotin (Vector Laboratories) solution, then lowered in the VTA (5.0–5.4 mm posterior to bregma, 0.5–1.0 mm lateral and 7.8–8.5 mm ventral). Signals were amplified tenfold (intracellular recording amplifier, Neurodata IR-283), filtered (cutoff 1 KHz, amplification tenfold, Cygnus Technologies Inc.), digitized (amplification tenfold, Axon Instruments Digidata 1322A) and acquired with Axoscope software (low-pass filter 5 KHz, high-pass filter 300 Hz, sampling 20 KHz). Baseline activity measurements were taken from the initial 5-min recording of the neuron, including mean firing rate, burst analysis, and duration and amplitude ratio of action potentials. Neurons with a mean baseline firing rate <6 Hz and a long-duration action potential (>1.5 ms) were considered to be putative dopamine neurons and were subjected to burst firing analysis based on established criteria¹⁸. Using these criteria, the majority of neurons recorded showed bursting activity (5/7, 71.42%). To assess the response to OFC stimulation, the mean value and s.d. of baseline activity was calculated using the 2,000 ms before the stimulation. Onset of inhibition (or excitation) was considered to be two consecutive bins after stimulation began in which the spike count was 2 s.d. or more below (or above) the mean bin value ($P < 0.001$). Offset of the response was considered to be two consecutive bins in which the bin values were no longer 2 s.d. from the mean bin value. When the value of 2 s.d. below the mean fell below zero, the number of consecutive bins required to signify the onset of inhibition was increased to maintain the same criterion for significance ($P < 0.001$).

Histology. Cells were labeled with Neurobiotin by passing positive current pulses (1.0–4.0 nA, 250-ms on/off square pulses, 2 Hz) and constant positive current (0.5–5.0 nA) through the recording electrode. For Neurobiotin and tyrosine hydroxylase immunohistochemistry, tissue was sectioned at 40 μm on a freezing microtome and collected in 0.1 M phosphate buffer in saline. After a 1-h pretreatment with 0.3% (vol/vol) Triton X-100 and 3% (vol/vol) normal goat serum in PBS, the sections were incubated overnight with Alexa 568-conjugated streptavidin (1:800, Molecular Probes) and a monoclonal mouse antibody to tyrosine hydroxylase (1:5,000, Swant). The sections were then rinsed in PBS several times

and incubated with a FITC-conjugated goat anti-mouse for 90 min (1:400, Jackson Laboratories). After rinsing in PBS, the sections were mounted on glass slides and coverslipped in Vectashield (Vector Laboratories), then examined under fluorescence on an Olympus FluoView 500 confocal microscope. Confocal images were captured in 2-μm optical steps. To confirm stimulating electrode placements, the OFC was sectioned at 50 μm and Nissl stained.

Computational modeling methods. *Task representation.* **Supplementary Figure 4a** shows the state-action diagram of the task in the intact model. Although simplified, this state-action sequence captures all of the key aspects of a trial. To account for errors, and in line with the behavioral data (**Fig. 1d**), we also included a 20% probability that the rat would make a mistake on a forced-choice trial—for example, going to the right reward port after a left signal.

Note that having the ‘enter left port’ state be the same on both forced and free trials allows our model to generalize between rewards received on free trials and those received on forced trials; that is, if a rat receives a long-delay reward by turning left on a forced trial, this architecture allows it to expect that the same long-delay reward will be delivered after turning left on a free trial.

After the rat moves to the reward port, it experiences state transitions according to one of two wait-for-reward state sequences (designated left row 1 to left row 3 or right row 1 to right row 3). These states indicate all the possible times that the reward could be delivered by the experimenter. Specifically, left row 1 is the state at which a reward drop is delivered to the left port on the ‘small’ and ‘short’ trials, and at which the first drop of reward is delivered on ‘big’ trials. Left row 2 is the time of the second reward drop on big trials. Rewards are never delivered at the wait left state, but this state, and the probabilistic self transition returning to wait left, implements a variable delay between the time of the early rewards and the long reward delivered at left row 3. Finally, after the left row 3 reward state, the task transitions into the end state, which signifies the end of the trial. State transitions are similar for the right-reward sequence.

Update equations. We modeled the task using an actor/critic architecture and a decaying eligibility trace⁴. We chose this framework because of its common use in modeling reinforcement learning in the brain; however, the same pattern of results was obtained when modeling state-action values (rather than state values) and the state-action-response-state-action (SARSA) temporal difference learning algorithm^{4,46,50}. Thus our results do not depend strongly on the particular implementation of TDRL.

At the start of the experiment, we set the critic’s initial values for each state, $V(s)$, and the actor’s initial preferences for action a , $B(s,a)$, to zero. At the beginning of each trial all eligibility traces $e(s)$ were set to zero, except that of the initial state s_1 , which was set to 1. The model dynamics were then as follows:

At each time step t of a trial, the model rat is in state s_t . First, the actor chooses action a_t with probability $\pi(s_t, a_t)$ given by

$$\pi(s_t, a_t) = \frac{\exp(B(s_t, a_t))}{\sum_a \exp(B(s_t, a))} \quad (1)$$

where $B(s_t, x)$ are the actor’s preference weights for action x in state s_t , and a enumerates all possible actions at state s_t . After taking action a_t , the model rat transitions to state s_{t+1} (whose value is $V(s_{t+1})$) and observes reward r_{t+1} . The prediction error, δ , is then computed according to

$$\delta = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \quad (2)$$

where $0 < \gamma < 1$ is the discount factor. At each time step, the prediction error signal is used to update all state values in the critic according to

$$V(s) \leftarrow V(s) + e(s)\eta\delta \quad (3)$$

where $0 < \eta < 1$ is the critic’s learning rate and $e(s)$ is the eligibility of state s for updating (see below). The actor’s preference for the chosen action a_t at state s_t is also updated according to the same prediction error:

$$B(s_t, a_t) \leftarrow B(s_t, a_t) + \beta\delta \quad (4)$$



where $0 < \beta < 1$ is the actor's learning rate. Finally, the eligibility trace of the last visited state, $e(s_{t+1})$, is set to 1 and all nonzero eligibility traces decay to 0 with rate $0 < \lambda < 1$ according to

$$\begin{aligned} e(s) &\leftarrow \lambda e(s) \quad \forall s \\ e(s_t) &= 1 \end{aligned} \quad (5)$$

The free parameters of the model (γ , η , β) were set manually to accord qualitatively with the data from the sham-lesioned rats and were not altered when modeling the OFC-lesioned group.

Value lesion (model 1). In this model, we removed all values from the critic: that is, we replaced all instances of $V(s)$ with 0 in the above equations. This resulted in prediction errors given by

$$\delta = r_{t+1} \quad (6)$$

which are only nonzero at the time of reward presentation.

Partial lesion within subjects (model 2). In this model, we assumed that f is the fraction of the critic that has been lesioned in each subject, leaving only $(1 - f)$ of the value intact. The prediction error was thus

$$\delta = r_{t+1} + (1 - f)\gamma(V_{M2}(s_{t+1}) - V_{M2}(s_t)) \quad (7)$$

and this prediction error was used as a teaching signal for the intact parts of the critic. Crucially, this led to learning of values $V_{M2}(s)$ in the intact parts of the critic, albeit at a slower rate than the unlesioned model. As a result of the error-correcting nature of TDRL, ultimately the learned values were

$$V_{M2}(s_t) = \frac{V(s_t)}{1 - f} \quad (8)$$

and the prediction errors at the end of training were similar to those of the unlesioned model, showing that the intact parts of the critic were able to fully compensate for the loss of parts of the critic, at least if training continued for enough trials.

Partial population lesion (model 3). In this model, we assumed that some fraction f of the population of rats had been successfully lesioned according to the value lesion model (model 1), whereas the other rats were unaffected.

The unlesioned fraction thus learn according to the prediction error of the unlesioned model, equation (2), whereas the lesioned fraction experience the prediction error of model 1, equation (6). Thus, averaged over the population, the prediction error is

$$\delta = r_{t+1} + (1 - f)\gamma(V_{M3}(s_{t+1}) - V_{M3}(s_t)) \quad (9)$$

Superficially, this prediction error resembles that of model 2. Unlike model 2, however, this prediction error is not directly used as a teaching signal; rather, each part of the population learns from a different prediction error signal (see above), and the learned values in the intact part of the population are simply $V_{M3}(s_t) = V(s_t)$.

OFC encoding of states (model 4). In this model we hypothesized that the effect of lesioning OFC is to change the state representation of the task. Specifically, we suggest that the OFC allows the rat to disambiguate states that require internal knowledge of impending actions but are otherwise externally similar, such as the state of "I have chosen the large reward in a free-choice trial" versus "I have chosen the small reward in a free-choice trial." We thus simulated the OFC lesion by assuming that the lesioned model no longer distinguishes between states based on the chosen action and cannot track correctly which type of reward is associated with which wait-for-reward state. Note that this ambiguity is caused not because the left and right reward ports are physically indistinguishable, but because, without knowledge of the mapping between the physical location of the ports and the abstract reward schedules at each port (in itself a type of internal 'expectancy' knowledge), the rat cannot tell which of the two possible wait-for-reward sequences it is in.

Mapping between prediction errors and VTA neural response. To facilitate comparison between the model results and those of the experiments, we transformed the prediction errors (which can be positive or negative) into predicted firing rates using a simple linear transformation,

$$\text{neural firing} = \text{baseline} + \text{scale factor} \times \text{prediction error} \quad (10)$$

with 5 spikes per second as the value for the baseline and negative prediction errors having a lower scale factor of 0.8 and positive prediction errors having a scale factor of 4.

50. Niv, Y., Daw, N.D. & Dayan, P. Choice values. *Nat. Neurosci.* **9**, 987–988 (2006).