SUPPLEMENTAL MATERIAL

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

Yuji K. Takahashi, Matthew R. Roesch, Robert C. Wilson, Kathy Toreson, Patricio O'Donnell, Yael Niv, and Geoffrey Schoenbaum

Effects of ipsilateral OFC lesions on activity of reward non-responsive dopamine neurons

We categorized 22 neurons and 26 neurons as reward non-responsive dopamine neurons in sham and OFC-lesion group, respectively (Fig. 2, main text). As a group, these neurons showed little evidence of error related activity, when they were analyzed using the same approach applied in the main text to the reward-responsive dopamine neurons (Fig. S1).

Effects of ipsilateral OFC lesions on activity of non-dopamine neurons

We categorized 429 neurons in sham group and 424 neurons in OFC-lesion group as non-dopamine neurons based on the cluster analysis (Fig. 2, main text). While some non-dopamine neurons did show phasic firing to reward, as a group they did not exhibit any characteristics of prediction error signaling in either sham or OFC-lesioned rats (Fig. S2).

Effects of ipsilateral OFC lesions on error signaling on delay trials

As noted in the main text, the lack of prediction error encoding in OFC-lesioned rats was also evident on delay trials. In the main text, this was examined when reward was delivered or omitted unexpectedly, which occurred at the transition from a long to short or short to long reward respectively. However another way to examine the effect of delay of reward is to examine changes in firing as the timing of delayed reward was titrated from 1 to 7 seconds. Importantly firing to this titrated, delayed reward was not included in the analysis of reward-evoked error signaling presented in the main text.

The timing of the delayed reward was titrated based on the rats' free-choice behavior. This was done to ensure that some responses would occur at each well to facilitate our neural analysis. One consequence of this manipulation was that the rats could not know precisely when to expect reward on these trials, since the timing of the delayed reward was not fixed. On the other hand, reward delivery on small reward trials was fixed and could be easily timed. The

difference in the ability the rats to time reward on these two trial types was evident in their licking behavior, which was constant during the 500 ms preceding the delayed reward but ramped up quickly in the 500 ms preceding the small reward (Fig. S3c and S3d, t-test; p's<0.01). There was no effect of ipsilateral OFC lesions on the change in licking (ANOVA; p=0.47). These unpredictable delayed rewards should elicit a positive prediction error, at least when compared to reward on the other trial types – particularly the trials in blocks 3^{sm} and 4^{sm} involving the small reward, which has the same low relative value as the delayed reward but a fixed and thus predictable time of delivery.

Consistent with this idea, in sham-lesioned rats, activity was significantly higher to the delayed, unpredictable reward than to the similarly-valued, but small, predictable reward (Fig. S3e, right). To quantify this effect, we computed the difference in firing for each neuron during the 500 ms following reward delivery on the two trial types (delayed – small). The distribution of these difference scores was significantly above zero, indicating that activity was typically higher for the less predictable reward (Fig. S3g right; Wilcoxon signed-rank test; p<0.01). This difference was markedly diminished in OFC-lesioned rats. This is evident in the population response (Fig. S3f right), which showed a much smaller average difference in firing to the delayed, unpredictable and small, predictable rewards than was observed in controls. Accordingly, the distribution of the difference scores comparing firing to the unpredictable delayed and predictable small reward was not shifted away from zero (Fig. S3h right; Wilcoxon signed-rank test; p=0.18), indicating that overall the population did not respond differently to reward on these two trial types, and the number of neurons that fired significantly more to delayed reward was no more than chance (6/50 neurons; X^2-test; p=0.12).

OFC lesions also disrupted suppression of firing in reward-responsive dopamine neurons on omission of the early, expected reward on these same delay trials (Fig. S3a and S3b). Rats in both groups expected reward to occur at this point, as indicated by the ramping increase in licking behavior preceding this unexpected omission (Fig. S3c and S3d); there was no effect of ipsilateral OFC lesions on the change in licking (t-test; p=0.31). However, while dopamine neurons in sham-lesioned rats suppressed firing upon reward omission (Fig. S3e, left), neurons in OFC-lesioned rats did not (Fig. S3d, left). To quantify this effect, we computed the difference in firing for each neuron in the 500 ms before and after reward omission (after minus before). The distribution of these scores was shifted significantly below zero in shams (Fig. S3g, left;

Wilcoxon signed-rank test; $p<0.01$), indicating that these neurons tended to fire less after reward than immediately before, whereas there was no shift in the distribution in OFC-lesioned rats (Fig. S3h, left; Wilcoxon signed-rank test; $p= 0.35$). In fact, the number of neurons that fired significantly less after reward omission in OFC-lesioned rats was no greater than chance and significantly less than the number in controls (14/30 *vs* 2/50 neurons; X^2-test; $p<0.01$).

We also examined the activity of reward-responsive dopamine neurons immediately before the delivery of the two reward types. In sham-lesioned rats, activity was significantly higher prior to delivery of the small, predictable reward than prior to delivery of the delayed, unpredictable reward (Fig. S3e, right). To quantify this effect, we computed the difference in firing for each neuron during the 500 ms before reward delivery on the two trial types (small – delayed). The distribution of these difference scores was shifted significantly above zero, indicating that activity was typically higher prior to delivery of a more predictable reward (Fig. S3g middle; Wilcoxon signed-rank test; $p<0.01$). Notably this difference was also evident in the reward-responsive dopamine neurons in OFC-lesioned rats. Thus the population activity was higher prior to delivery of the small, predictable reward (Fig. S3f, right), and the distribution of difference scores was significantly shifted above zero (Fig. S3h, middle; Wilcoxon signed-rank test; $p<0.01$). There was no significant difference in the distribution of these difference scores between sham and OFC-lesioned rats (Mann Whitney U test; $p=0.53$).

Additional discussion of modeling results

It is important to note that the models we consider are all qualitatively different from each other. In model 2 the learning rate is slower than the non-lesioned model, as part of the critic is lesioned. However this does not prevent asymptotically correct values from being learned for the two conditions (low and high reward port) as well as for the two choices in the free-choice trials. Thus prediction errors to reward should decrease with training, and prediction errors to cues/choices should increase with training, predictions which are both at odds with the actual data. On the other hand, in model 3 only part of the population is fully lesioned (no learning of values) and part is not lesioned at all (intact learning). This results in averaged results that are a mixture of these two cases, which again does not match up with the empirical data. Importantly, in both cases, the qualitative result is not dependent on the specific choice of

learning rate for the implementation. In model 4, however, there is a lesion in the state representation such that correct values for the two states of receiving a high and low reward cannot be fully learned, and values for the two choices in the free-choice trials are not learned at all, no matter how long learning goes on, and no matter what learning rate parameter is chosen.

With regard to Model 4, we note that there are other possible simplifications of the state space, however, those would not be consistent with the empirical data. For instance, OFC lesions might also eliminate the critic's access to state information related to the trial type (i.e., odor cue), further simplifying the state space in Figure 6, but this would predict that firing rates on all forced trials should be identical. In this regard, it is worth noting that what the data show, and the model attempts to reproduce, is what OFC contributes to error coding in this specific task that no other area can provide. Thus the fact that information about the odor cue does not seem to require OFC should not be taken as evidence that OFC does not represent that component of the state space, only that information about such external cues in our particular task is also provided by other neural systems.

SARSA and Q-Learning instantiations of Model 4:

Our results are true for a variety of temporal difference RL methods, and not dependent on the actor-critic architecture. To demonstrate this, we also instantiated SARSA and Q-Learning versions of model 4. Rather than separately learning state values and an action policy, these algorithms use temporal difference learning to directly learn the value of taking particular actions and base their decisions on these learned action values.

Both of these algorithms learn the value $Q(s_t, a_t)$ of taking action a_t in state s_t via a temporal difference update with learning rate α

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\delta \tag{10}$$

Where the prediction error, δ, is computed according to

$$\delta = r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \tag{11}$$

for SARSA and

$$\delta = r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \tag{12}$$

for Q-Learning. This means that SARSA evaluates prediction errors based on the difference in values of the current action and the previous one ("on-policy" learning), while Q-learning compares the best available action to the previous one, regardless of which is the currently chosen action ("off-policy" learning; for more details see ref. [Sutton & Barto]).

Actions are selected using a softmax policy such that the probability of choosing action a_t when in state s_t is given by

$$\pi(s_t, a_t) = \frac{\exp(\beta Q(s_t, a_t))}{\sum_b \exp(\beta Q(s_t, b))} \qquad (13)$$

For some inverse temperature parameter β.

The state-action spaces for these models are shown in figure S5a for the sham and lesioned models, and are the same as for the actor-critic version of model 4 except for the addition of the small black circles denoting the actions whose values are learned.

In figure S5b we show the results of simulating the SARSA model. As with the actor-critic model, this model is able to capture the full range of results seen in the data.

In figure S5c we show the results of simulating the Q-learning model. Although this model predicts the correct pattern of prediction errors in the lesioned case, its prediction for the "free" odor case in sham-lesioned rats is at odds with the data (red arrow). This is because the Q-learning prediction error involves the best available action (not necessarily the chosen one) and thus at the time of decision for the free choice odor the prediction error does not differentiate between trials in which the high or the low option were chosen:

$$\begin{aligned} \delta &= \gamma \max_a Q(\text{free}, a) - Q(\text{odor}, \text{sniff}) \\ &= \gamma Q(\text{free}, \text{high}) - Q(\text{odor}, \text{sniff}) \end{aligned} \qquad (14)$$

regardless of the decision actually made. This suggests, similar to previous results from non-lesioned animals [ref to Morris et al.] that the SARSA algorithm better accounts for animal learning than the Q-learning algorithm.

FIGURE LEGENDS

Figure S1: Changes in activity of reward non-responsive dopamine neurons (Fig. 2, main text) in response to unexpected reward delivery and omission. (a and b) Heat plots show average response of all reward non-responsive dopamine neurons in sham **(a)** and OFC-lesioned rats **(b)** to introduction of unexpected reward in block 2^{sh} (upper plots, black arrows) and omission of expected reward in block 2^{lo} (lower plots, grey arrows). Activity in each plot is synchronized to time of reward (or omission). **(c and f)** Distribution of difference scores comparing activity to an unexpected reward early versus late in sham **(c)** and OFC-lesioned rats **(f)**. Difference scores were computed from the average firing rate of each neuron in the first 5 minus the last 15 trials during the 500ms after delivery of an unexpected reward (blocks 2^{sh}, 3^{bg}, and 4^{bg}). 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 test (p) and the average index (u). **(d and g)** Distribution of difference scores comparing activity to omission of an expected reward early versus late in sham **(d)** and OFC-lesioned rats **(g)**. Difference scores were computed from the average firing rate of each neuron in the first 5 minus the last 15 trials during the 500ms after omission of an expected reward (blocks 2^{lo} and 4^{sm}). 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 test (p) and the average index (u). **(e and h)** Scatter plots show the relationship between difference scores for individual neurons to unexpected reward and reward omission in sham **(e)** and OFC-lesioned rats **(h)**.

Figure S2: Changes in activity of non-dopamine neurons (Fig. 2, main text) in response to unexpected reward delivery and omission. (a and b) Distribution of baseline firing rates for non-dopamine neurons in sham **(a)** and OFC-lesioned rats **(b)**. **(c)** Average baseline firing rates for non-dopamine neurons. S, sham; L, lesioned. Asterisks indicate planned comparisons revealing statistically significant differences as described in the text (t-test or other comparison, $p < 0.05$ or better); 'ns' denotes non-significant. Error bars, SEM. Heat plots show average response of all non-dopamine neurons in sham **(d)** and OFC-lesioned rats **(e)** to introduction of unexpected reward in block 2^{sh} (upper plots, black arrows) and omission of expected reward in block 2^{lo} (lower plots, grey arrows). Activity in each plot is synchronized to time of reward (or omission). **(f and i)** Distribution of difference scores comparing activity to an unexpected reward early versus late in sham **(f)** and OFC-lesioned rats **(i)**. Difference scores were computed

from the average firing rate of each neuron in the first 5 minus the last 15 trials during the 500ms after delivery of an unexpected reward (blocks 2^{sh}, 3^{bg}, and 4^{bg}). 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 test (p) and the average index (u). **(g and j)** Distribution of difference scores comparing activity to omission of an expected reward early versus late in sham **(g)** and OFC-lesioned rats **(j)**. Difference scores were computed from the average firing rate of each neuron in the first 5 minus the last 15 trials during the 500ms after omission of an expected reward (blocks 2^{lo} and 4^{sm}). 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 test (p) and the average index (u). **(h and k)** Scatter plots show the relationship between difference scores for individual neurons to unexpected reward and reward omission in sham **(h)** and OFC-lesioned rats **(k)**.

Figure S3: Activity of reward-responsive dopamine neurons (Fig. 2, main text) in response to and unpredictable delayed reward. (a and b) Line deflections indicate the time course of well entry and reward omission and delivery on delay and small reward trials. Dashed lines show when reward is omitted (long delay trials) and solid lines show when reward is delivered (delay and small reward trials). **(c and d)**. Licking aligned on omission (left) and delivery of reward (right) on delay (blue) and small (red) reward. Licking increased significantly before delivery of reward on small trials and before omission of reward on delay trials; licking did not change significantly prior to reward delivery on delay trials. Inset bar graphs indicate slope of rise in licking behavior during the 500 ms preceding reward omission (left) and delivery on delay (blue in right) and small (red in right) reward trials (*$p<0.01$ or better; t-test). Error bars = SEM. **(e and f)** Average firing rate of reward-responsive dopamine neurons in sham **(e)** and OFC-lesioned rats **(f)** aligned on omission (left) and delivery of reward (right) on delay (blue) and small (red) reward trials. **(g and h)** Distributions of the difference scores in sham **(g)** and OFC-lesioned rats **(h)**. 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 test (p) and the average index (u). Leftmost panels indicate the distribution of the difference in firing rate during reward delivery between delay and small reward trials (delay minus small). The distribution of these scores was significantly shifted above zero in sham (Wilcoxon; $p<0.01$) but not in OFC-lesioned rats (Wilcoxon; $p=0.18$). In addition, the numbers of neurons that fired

significantly more to delayed reward was no more than chance in OFC-lesioned group (6/50 neurons; X^2-test; p=0.12). Middle panels indicate the distribution of the difference in firing for each neuron in the 500ms before reward delivery between small and delayed reward trials (small minus delay). The distribution of these scores was significantly shifted above zero in both sham (Wilcoxon; p<0.01) and OFC-lesioned rats (Wilcoxon; p<0.01). Right panels indicate the distribution of the difference in firing for each neuron in the 500ms before and after reward omission (after minus before). The distribution of these scores was significantly shifted below zero in sham group (Wilcoxon; p<0.01), but not in OFC-lesioned group (Wilcoxon; p=0,35). The proportion of neurons that fired significantly less after reward omission in OFC-lesioned group was significantly less than in sham (14/30 vs 2/50 neurons; X^2-test; p<0.01).

Figure S4: State spaces used in the model and SARSA and Q learning algorithms. (a) Schematics depict the state spaces used for the modeling. Top figure shows the full state space. Bottom figure shows the state space used to model the OFC lesion in Model 4. **(b-c)** Results from Model 4 implemented using SARSA **(b)** or Q learning **(c)** algorithms.

Fig.S1

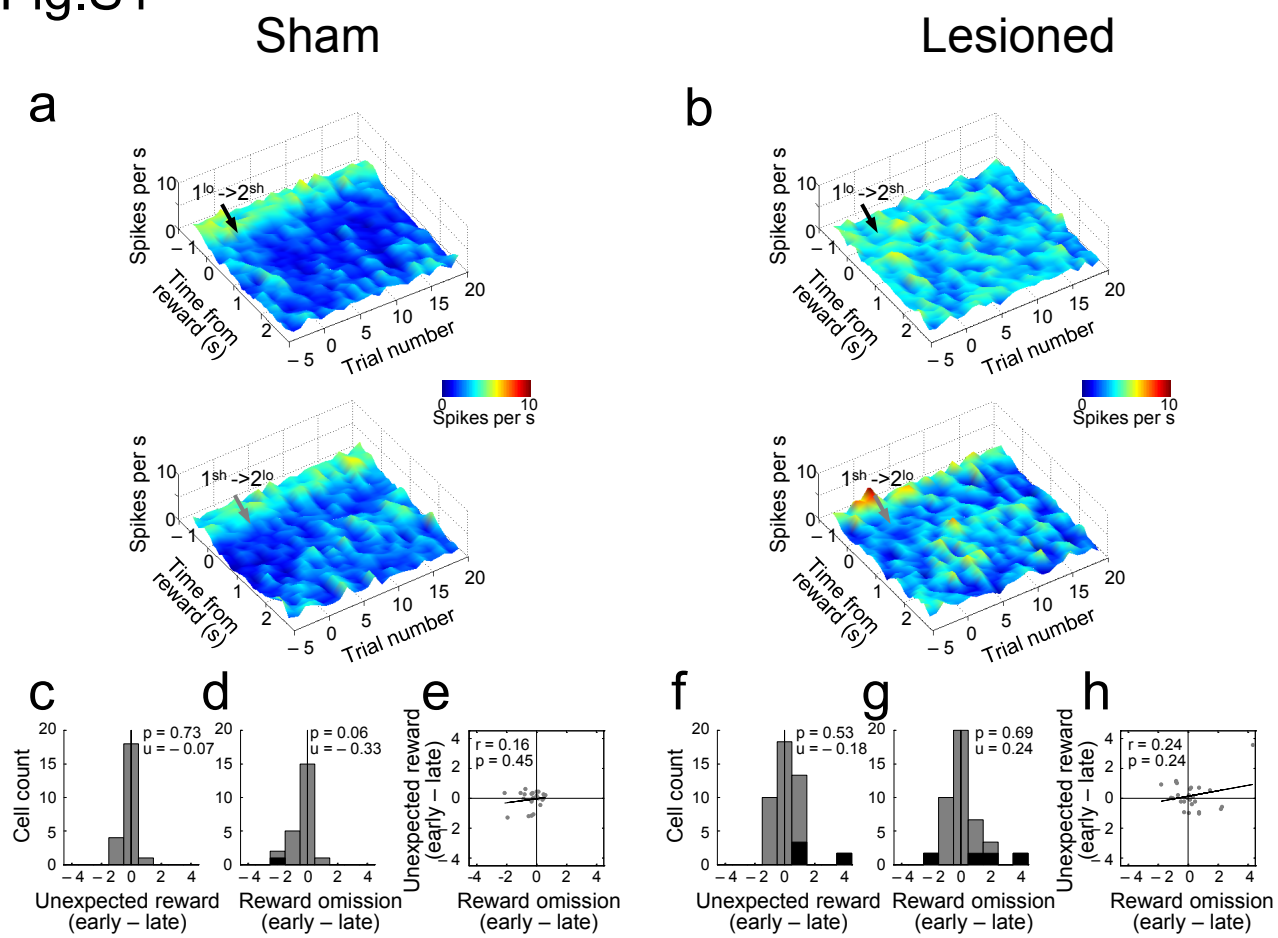

Sham
Lesioned
a
b
Spikes per s
Time from reward (s)
Trial number
1lo ->2sh
1sh ->2lo
c
d
e
f
g
h
Cell count
p = 0.73
u = – 0.07
p = 0.06
u = – 0.33
p = 0.53
u = – 0.18
p = 0.69
u = 0.24
r = 0.16
p = 0.45
r = 0.24
p = 0.24
Unexpected reward (early – late)
Reward omission (early – late)

Fig.S2

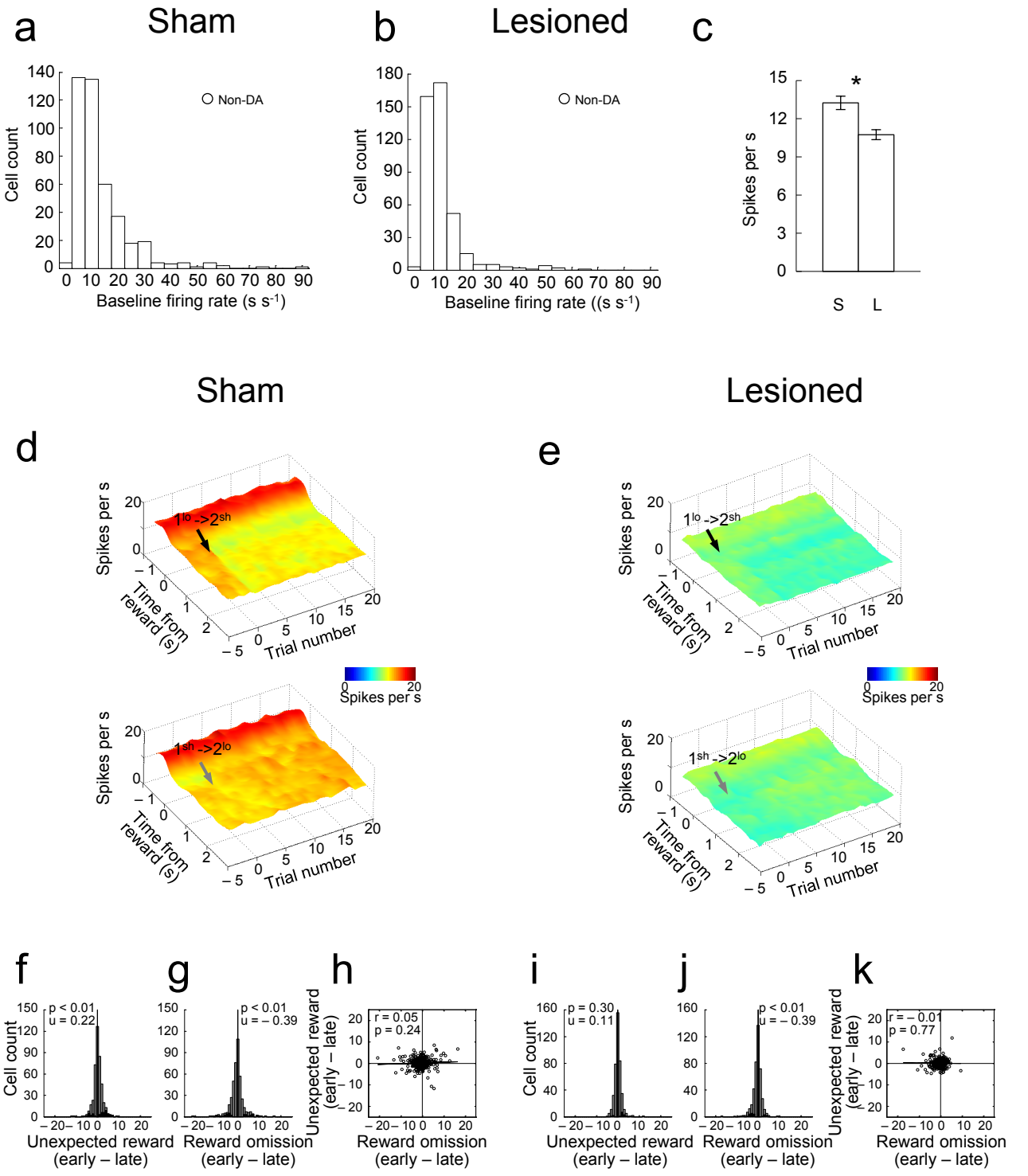

a
Sham
Non-DA
Cell count
Baseline firing rate (s s-1)
b
Lesioned
Non-DA
Cell count
Baseline firing rate ((s s-1)
c
Spikes per s
S
L
Sham
Lesioned
d
e
Spikes per s
Time from reward (s)
Trial number
1lo ->2sh
1sh ->2lo
f
g
h
i
j
k
Cell count
p < 0.01
u = 0.22
p < 0.01
u = – 0.39
p = 0.30
u = 0.11
r = 0.05
p = 0.24
r = – 0.01
p = 0.77
Unexpected reward (early – late)
Reward omission (early – late)

Fig.S3

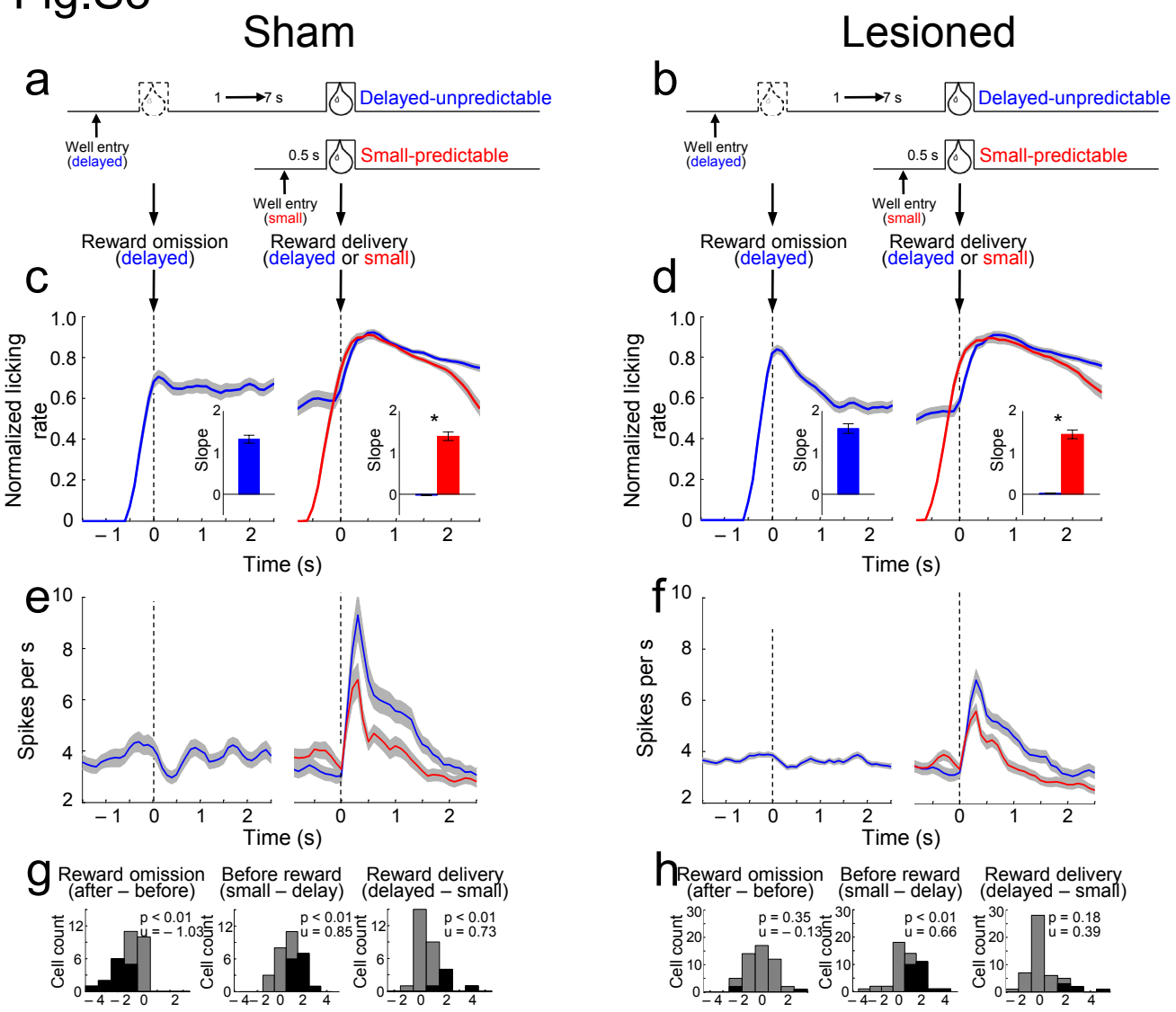
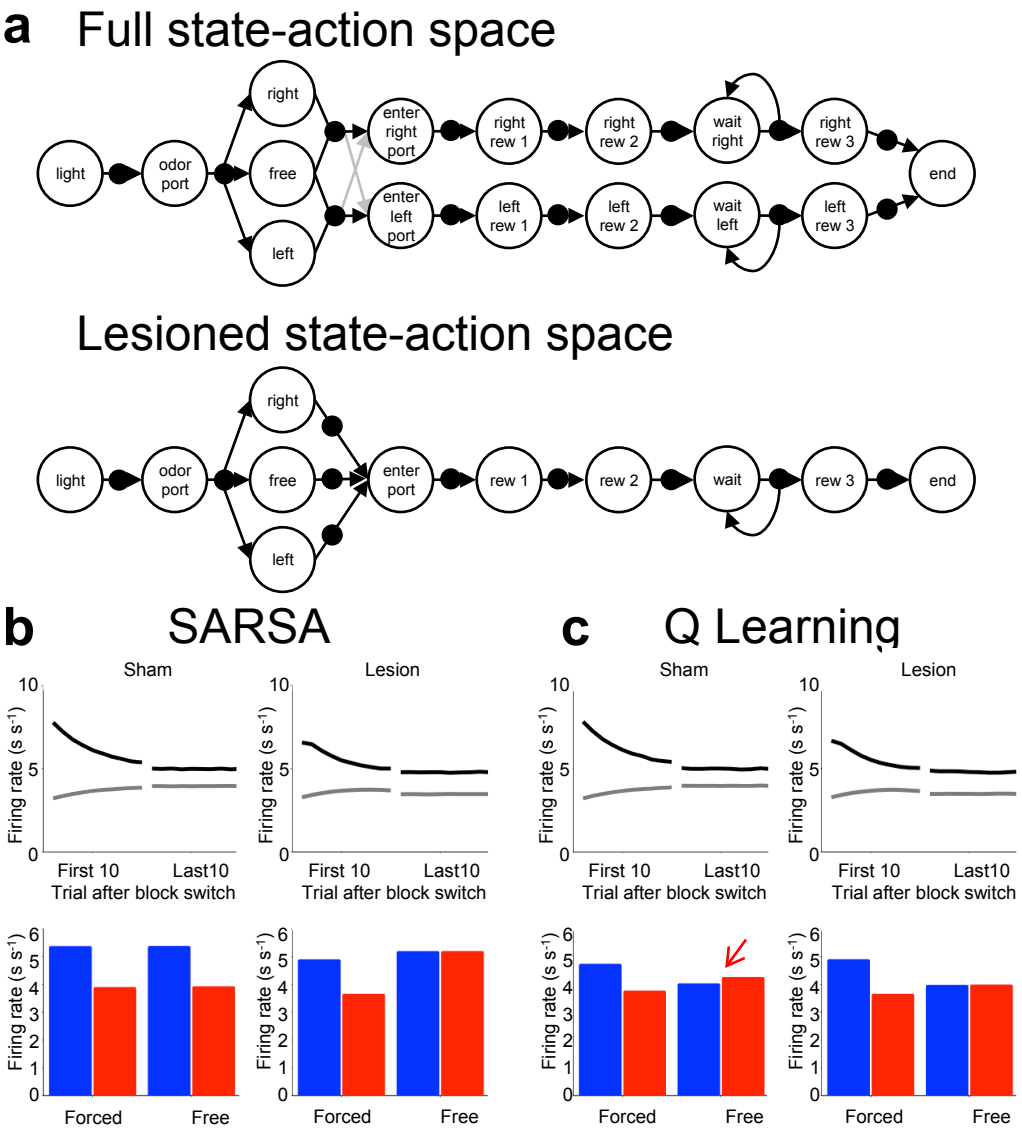

Sham
Lesioned
a
b
1 → 7 s
Delayed-unpredictable
Well entry (delayed)
0.5 s
Small-predictable
Well entry (small)
Reward omission (delayed)
Reward delivery (delayed or small)
c
d
Normalized licking rate
Slope
Time (s)
e
f
Spikes per s
g
h
Reward omission (after – before)
Before reward (small – delay)
Reward delivery (delayed – small)
Cell count
p < 0.01
u = – 1.03
p < 0.01
u = 0.85
p < 0.01
u = 0.73
p = 0.35
u = – 0.13
p < 0.01
u = 0.66
p = 0.18
u = 0.39

Fig.S4


a
Full state-action space
light
odor port
right
free
left
enter right port
enter left port
right rew 1
right rew 2
wait right
right rew 3
left rew 1
left rew 2
wait left
left rew 3
end
Lesioned state-action space
light
odor port
right
free
left
enter port
rew 1
rew 2
wait
rew 3
end
b
SARSA
c
Q Learning
Sham
Lesion
Firing rate (s s-1)
First 10
Last10
Trial after block switch
Forced
Free