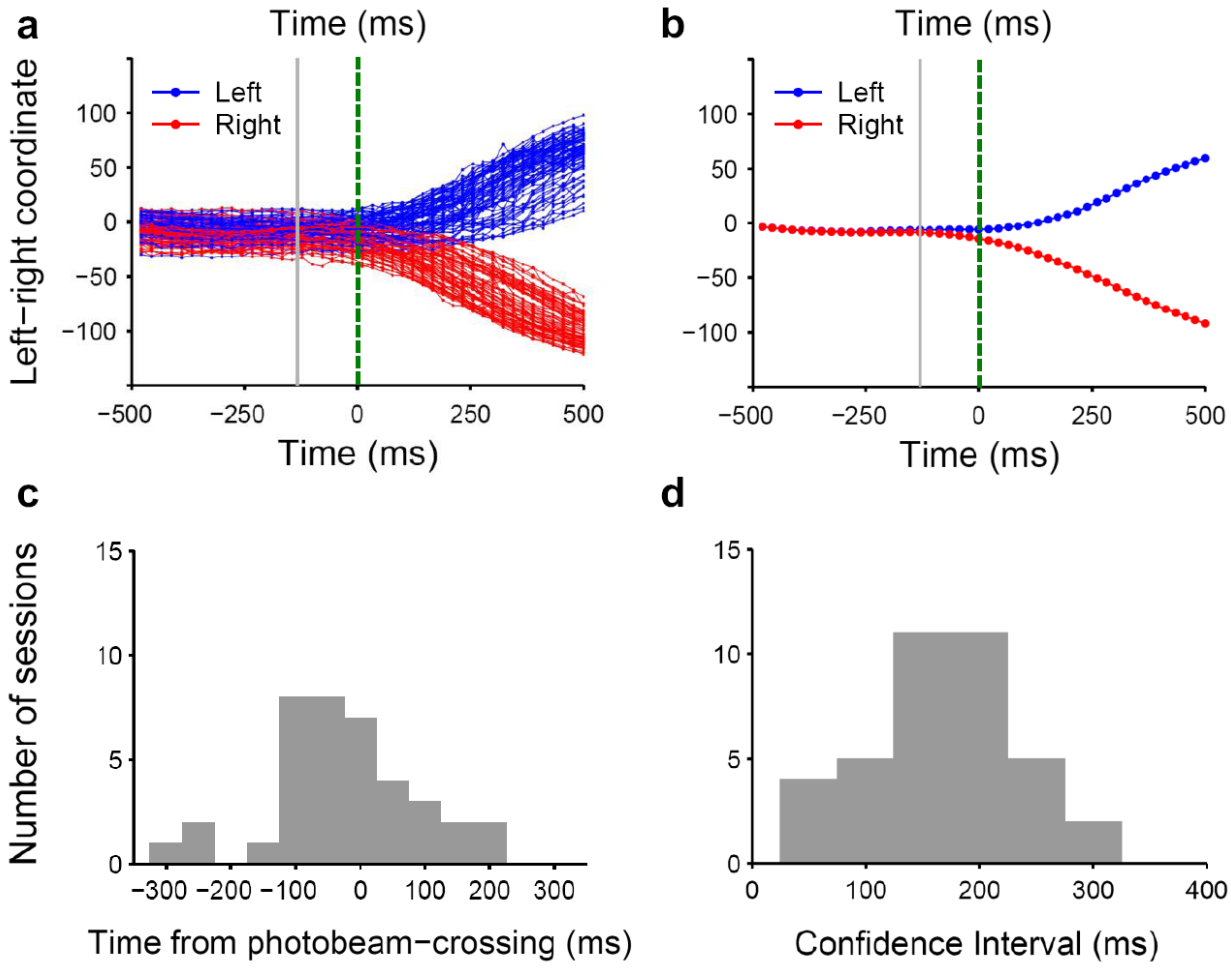
SUPPLEMENTARY INFORMATION

Sul JH, Jo S, Lee D & Jung MW

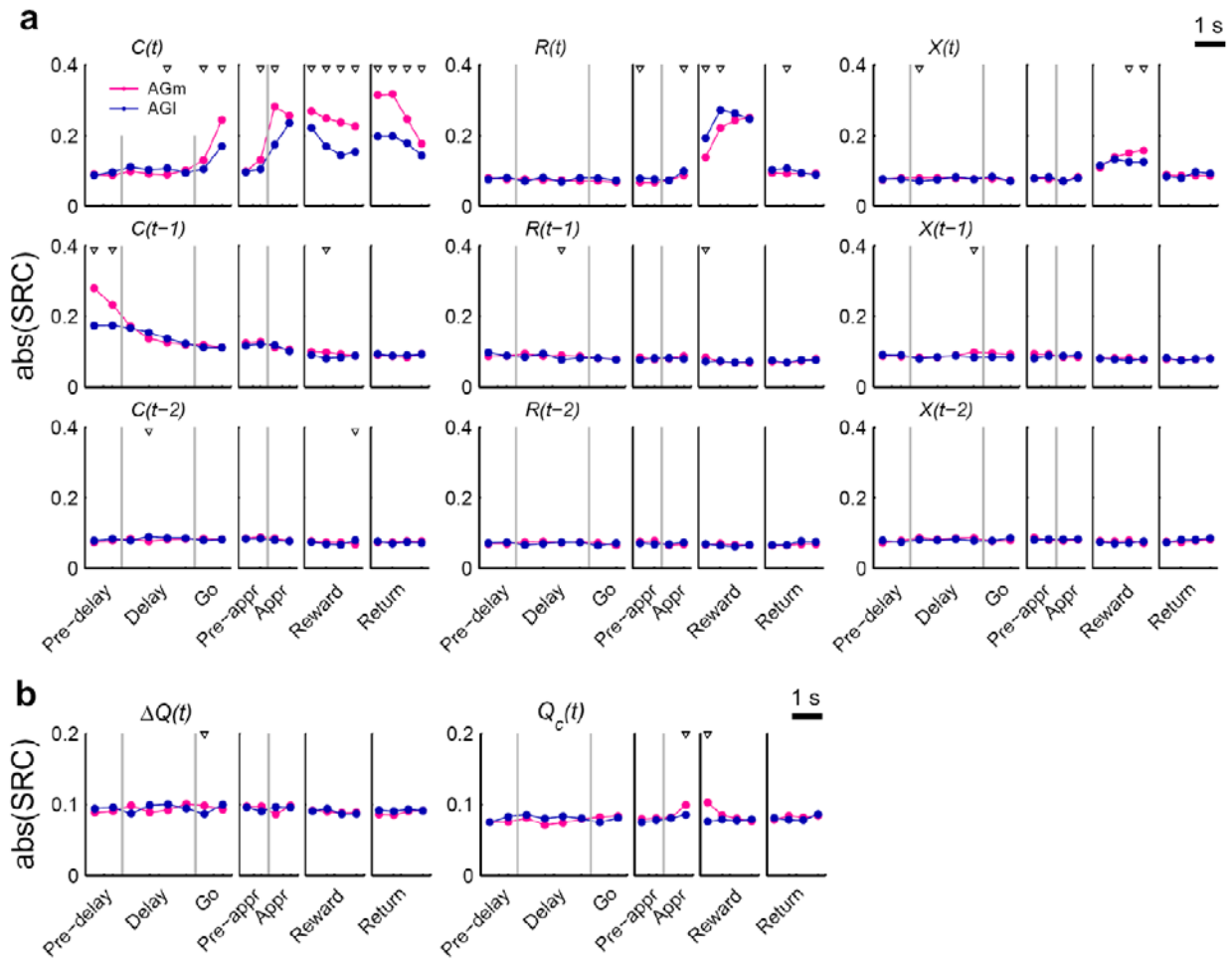
Role of rodent secondary motor cortex in value-based action selection

SUPPLEMENTARY FIGURES

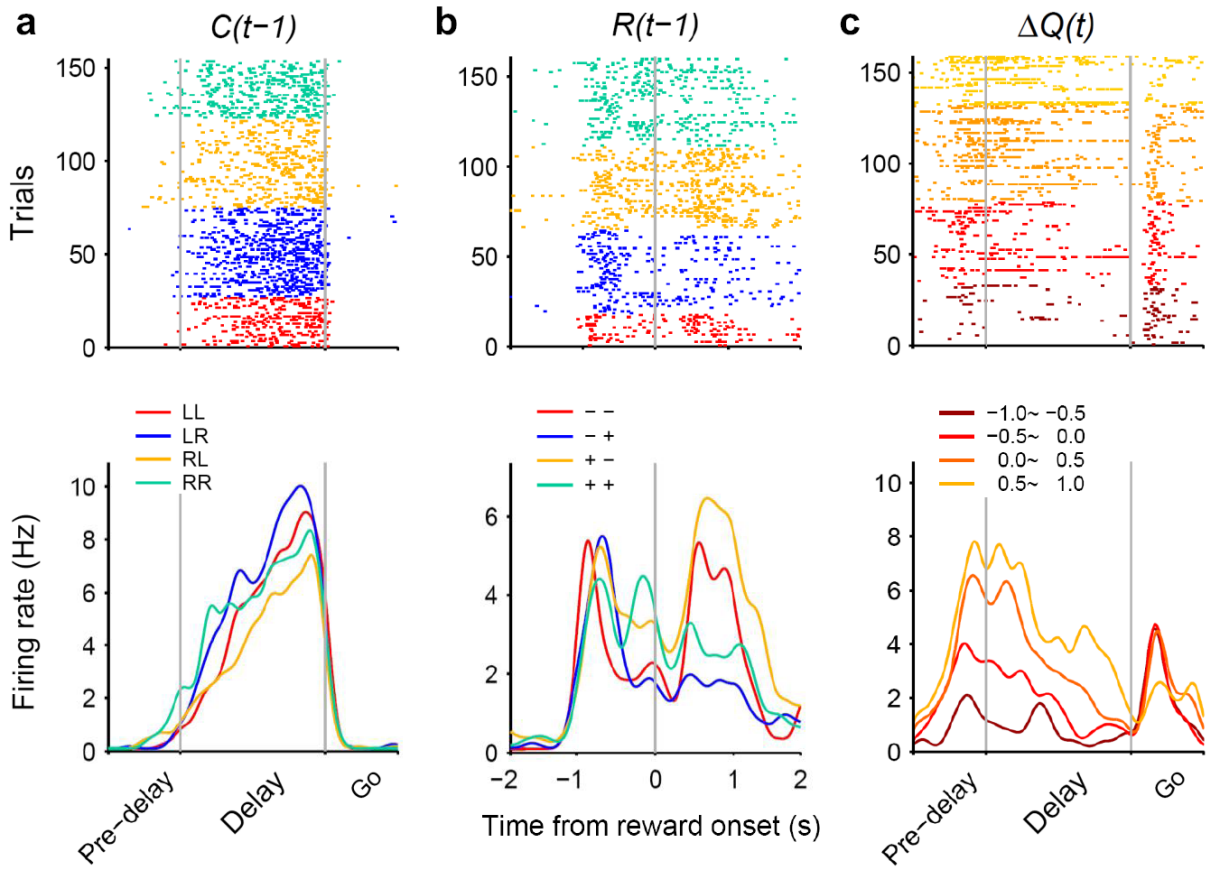


Supplementary Figure 1. Method used to determine the onset of approach stage. (**a**-**b**) The graphs show the time course of left-right coordinates of the animal's movement trajectories (sampling interval=16.7 ms) near the onset of the approach stage during an example recording session (**a**, individual trials; **b**, mean). Blue and red colors indicate trials associated with the left

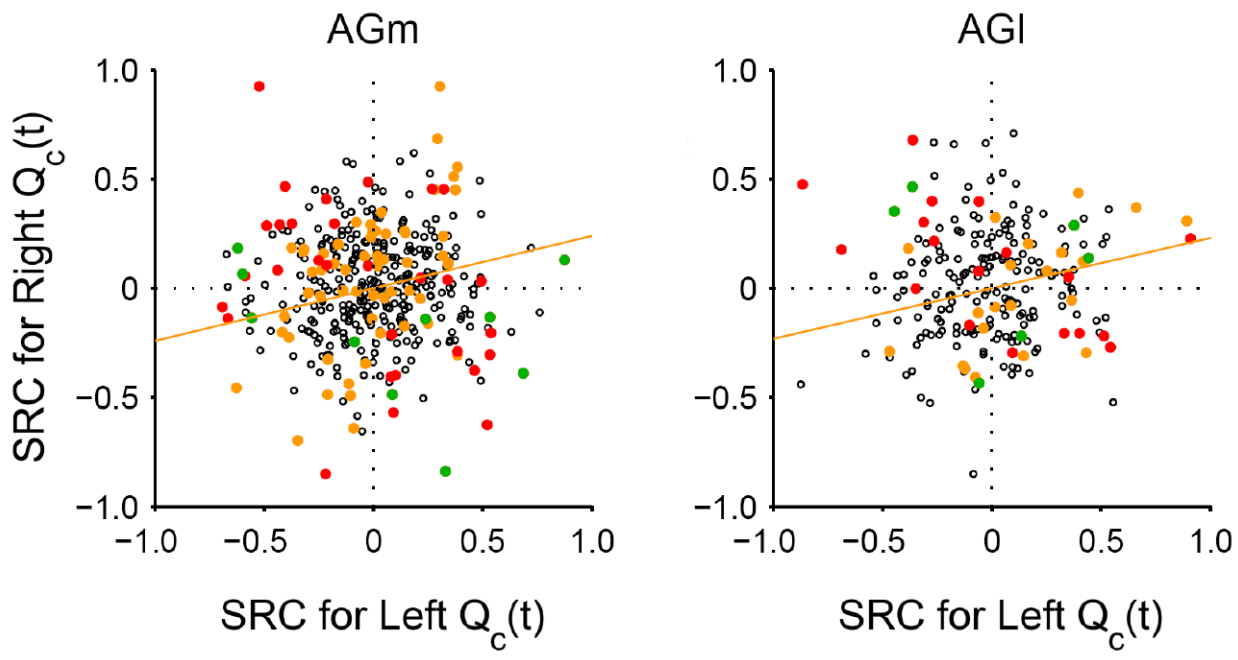
and right goal choice, respectively. Green dotted line (0 ms) indicates the time when the animal broke the photobeam that was placed near the upper branching point, and the gray line (approach onset) indicates the time when the left-right positions became significantly different for the left- and right-choice trials (*t*-test, $p < 0.05$) for the first time within ± 0.5 s time window. (**c**) Distribution of the approach onset across all sessions. Time 0 is when the animal broke the photobeam (green dotted line in **a** and **b**) with the positive numbers indicating the divergence of trajectory after breaking the photobeam. (**d**) Distribution of 95% confidence interval for the time of approach onset for each session. Left- and right-choice trials were randomly sampled with replacement from the original left- and right-choice trials, respectively (same numbers as the original left- and right-choice trials), and the time of approach onset was determined in the same way as for the original data. This procedure was repeated 1,000 times, and 95% confidence interval for the time of approach onset was determined based on their distribution for each session.



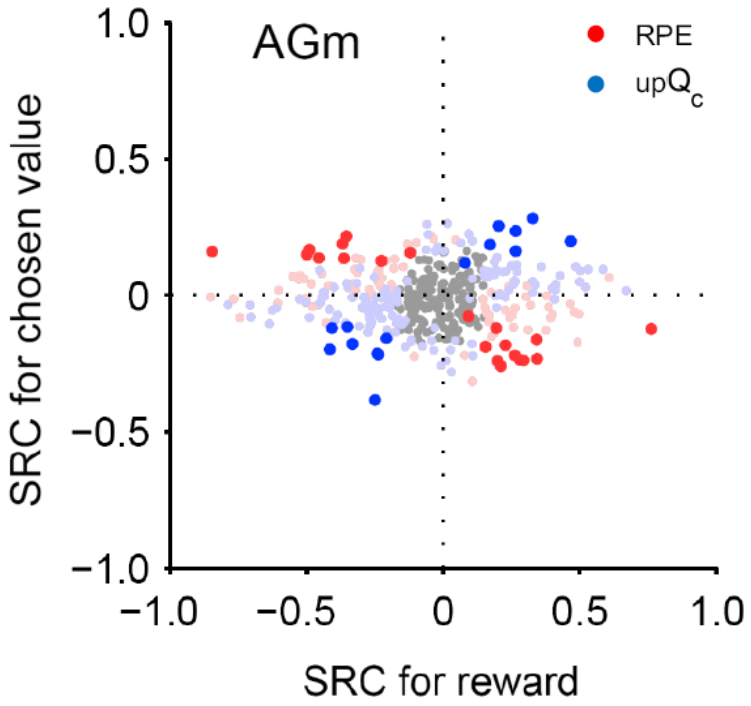
Supplementary Figure 2. Effect size of neural signals related to the animal's choice, reward, and values. The graphs show absolute values of the standardized regression coefficients [abs(SRC)] obtained with the same regression models used in Fig. 3a (**a**, *eq. 1*) and 4a (**b**, *eq. 2*). Same format as in Fig. 3a and 4a. Triangles indicate significant differences across regions (*t*-test, $p < 0.05$).



Supplementary Figure 3. Examples of AGm neurons encoding previous choice, previous reward, or decision value. (**a**) An example AGm neuron encoding the animal's choice in the current and previous trial. Trials were grouped according to the sequence of choices in the previous and current trial (L, left; R, right; e.g., LR, left and right goal choice in the previous and current trial, respectively). (**b**) An example AGm neuron encoding reward in the current and previous trial. Trials were grouped according to the sequence of reward delivery in the previous and current trial (+, rewarded; -, unrewarded; e.g., +-, rewarded and unrewarded in the previous and current trial, respectively). (**c**) An example AGm neuron encoding decision value. Both a spike raster (top) and spike density functions (bottom; Gaussian kernel, σ=100 ms) are shown for each example.



Supplementary Figure 4. Relationship between standardized regression coefficients for chosen value in the left and right goal-choice trials. Trials were divided according to the animal's goal choice (left vs. right) and neural activity during the first 1 s of the reward stage was analyzed with the following regression model: $S(t) = a_0 + a_1R(t) + a_2\Delta Q(t) + a_3Q_c(t) + a_4C(t\text{-}1) + a_5R(t\text{-}1) + A(t) + \varepsilon(t)$, where $A(t)$ is the autoregressive term (see Online Methods, *eq. 2*). The abscissa and ordinate indicate regression coefficients for the left- and right-choice trials, respectively. Orange circles indicate those neurons encoding chosen value (*eq. 2*). Red circles indicate those neurons showing significant chosen action×chosen value interaction [$C(t)\times Q_c(t)$] as determined by the following regression model: $S(t) = a_0 + a_1C(t) + a_2R(t) + a_3X(t) + a_4Q_L(t) + a_5Q_R(t) + a_6Q_c(t) + a_7C(t)\times Q_c(t) + a_8C(t\text{-}1) + a_9R(t\text{-}1) + A(t) + \varepsilon(t)$. Green circles denote those neurons encoding both chosen value (*eq. 2*) and showing significant chosen action×chosen value interaction (above regression). The neurons encoding chosen value (*eq. 2*) were used to determine the best-fitting lines (orange lines) and to calculate the correlation coefficients (AGm: r=0.223, p=0.068; AGl: r=0.171, p=0.172).



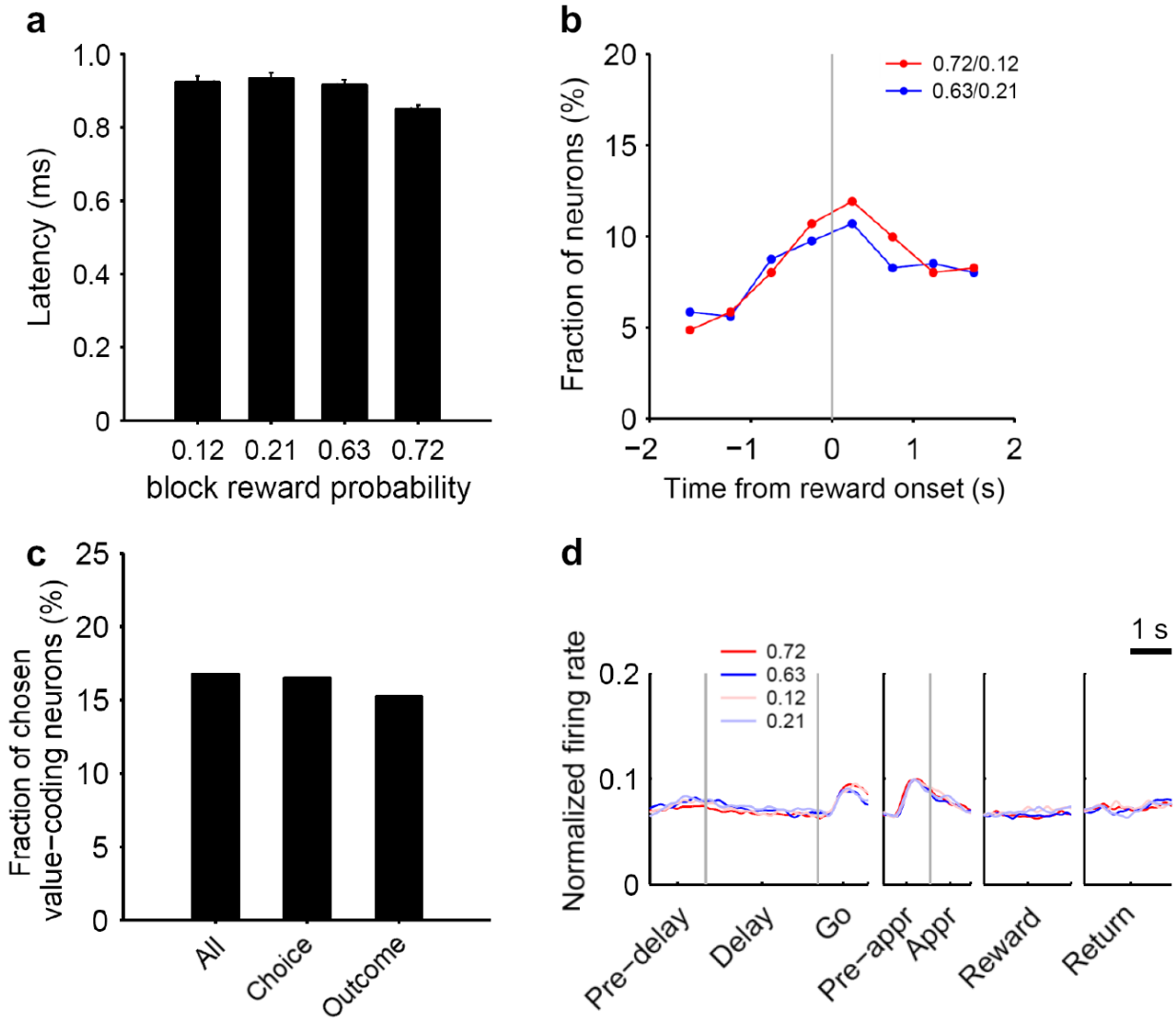
Supplementary Figure 5. Relationship between the coefficients for chosen value and reward. Standardized regression coefficients (SRC) for the current reward [*R(t)*] were plotted against those for the chosen value [$Q_c(t)$] for neural activity in the AGm during the first 1 s of the reward stage. Each dot corresponds to one neuron. Saturated colors indicate those AGm neurons that encoded both reward and chosen value, and light colors indicate those that encoded either reward or chosen value only. The remaining neurons are shown in gray. Neural signals for chosen value and reward can be combined to compute RPE or to update chosen value, although we cannot exclude the possibility that they are combined to compute other unknown variables. Therefore, we compared whether RPE or updated chosen value [$upQ_c(t)$] better explained activity of those AGm neurons that encoded both chosen value and reward during the first 1 s of the reward stage. Specifically, we examined which of the following models better accounted for neuronal activity:

$$S(t) = a_0 + a_1C(t) + a_2\Delta Q(t) + a_3C(t\text{-}1) + a_4R(t\text{-}1) + a_5RPE + A(t) + \varepsilon(t) \quad (eq.\ 3)$$

$$S(t) = a_0 + a_1C(t) + a_2\Delta Q(t) + a_3C(t\text{-}1) + a_4R(t\text{-}1) + a_5upQ_c(t) + A(t) + \varepsilon(t) \quad (eq.\ 4)$$

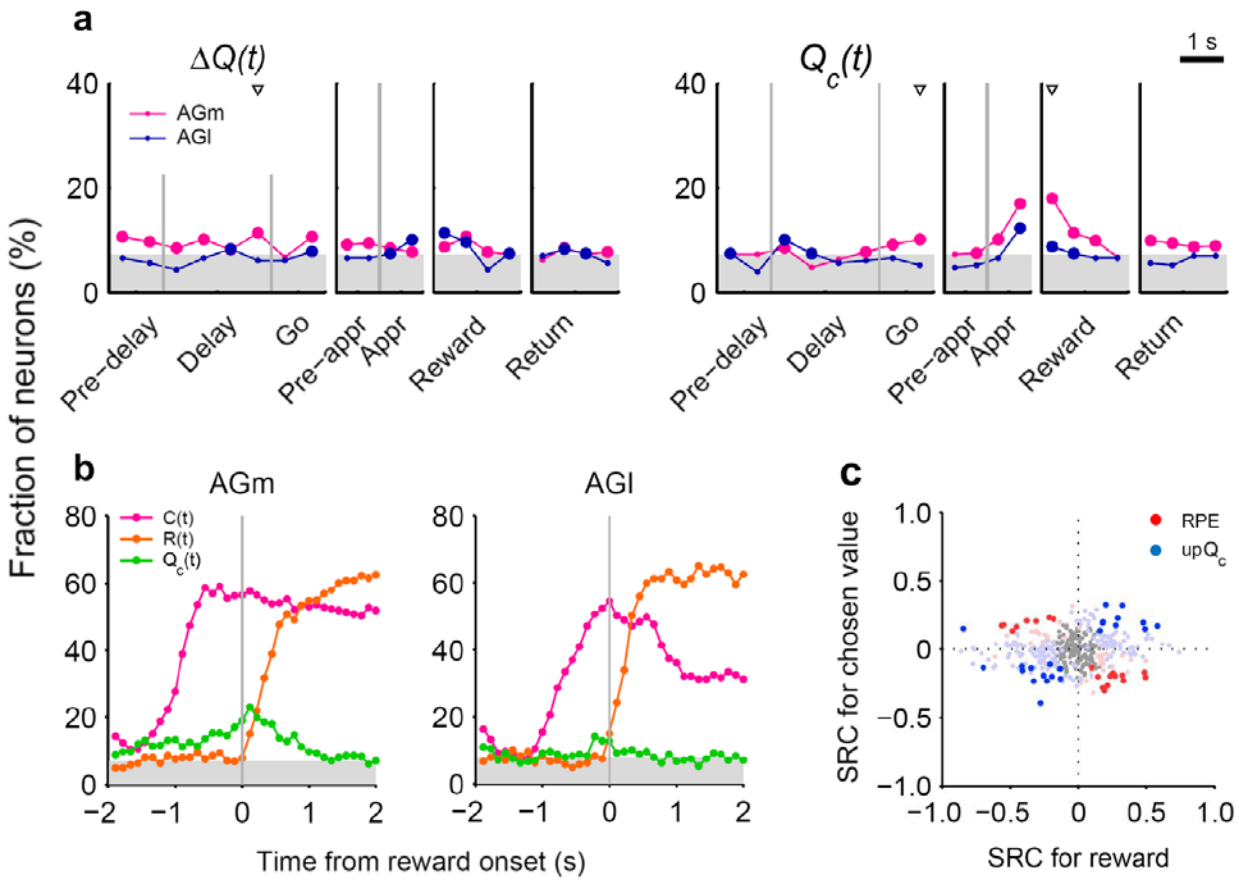
where $RPE = R(t) - Q_c(t)$, $upQ_c(t) = Q_c(t) + \alpha RPE$, α is the learning rate estimated from the RL model, and *A(t)* is the autoregressive term (see Online Methods, *eq. 2*). Because RPE is the difference between the reward and chosen value, RPE-coding neurons are expected to have

opposite signs of the coefficients for reward [$R(t)$] and chosen value [$Q_c(t)$]. On the other hand, because updated chosen value is determined by the weighted sum of these variables [updated chosen value = $\alpha R(t) + (1-\alpha)Q_c(t)$, where α is learning rate and $0 < \alpha < 1$], updated chosen value-coding neurons are expected to have same signs for their coefficients. Among 36 AGm neurons that significantly modulated their activity according to both reward and chosen value (*eq. 2*; saturated colors), 21 and 15 had opposite and same signs of the coefficients for reward and chosen value, respectively. As expected, activity of the former (opposite signs) was better accounted for by the regression model containing RPE (*eq. 3*, red colors), whereas activity of the latter (same signs) was better accounted by the regression model containing updated chosen value (*eq. 4*, blue colors). These results suggest that reward and chosen value signals are combined to compute RPE as well as to update the value of chosen action in the AGm.

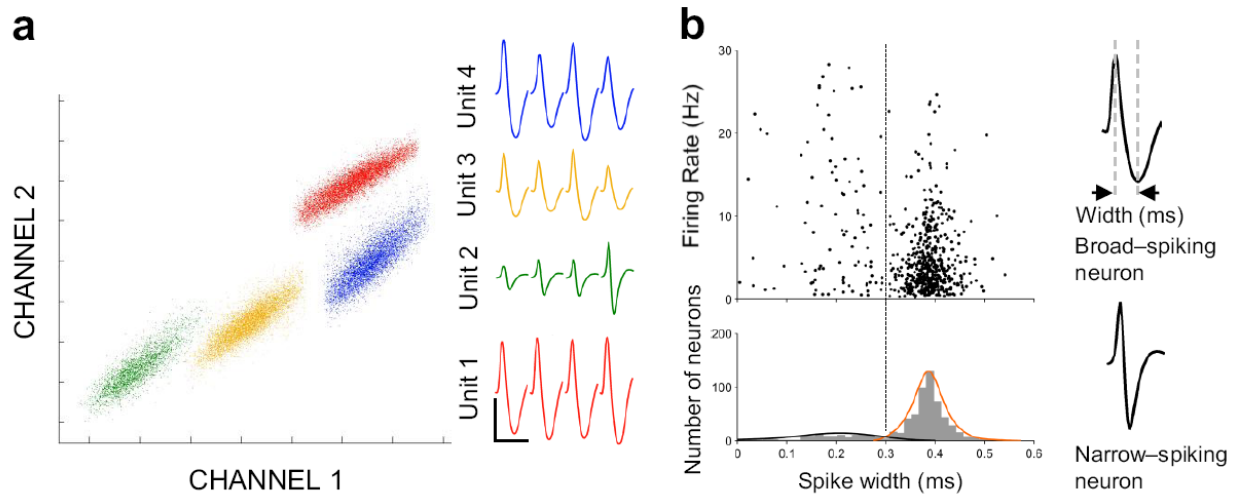


Supplementary Figure 6. Evidence against the possibility that chosen value signals in the AGm merely reflect motivation-regulated movement preparation. (**a**) The duration of the approach stage (mean±SEM) as a function of the block reward probability of the chosen goal. The duration of the approach stage was significantly shorter only when approaching the goal with the reward probability of 0.72 (one-way ANOVA, $F_{(3,6085)}$=9.381; $p < 0.001$; post-hoc Bonferroni pair-wise comparisons, 0.72 vs. 0.12, p=0.002; 0.72 vs. 0.21, $p < 0.001$; 0.72 vs. 0.63, $p < 0.001$; 0.63 vs. 0.12, p=1.000; 0.63 vs. 0.21, p=1.000; 0.21 vs. 0.12, p=1.000). (**b**) Nevertheless, the time course of chosen value signals in the AGm was similar when the same

analysis was repeated after dividing the data into two halves (0.72/0.12 vs. 0.63/0.21 blocks). Significant difference was found in none of the time intervals (χ^2-test, $p > 0.05$ for all time intervals). Thus, whereas the animals might have been more motivated when choosing the 0.72 probability goal, this difference cannot account for chosen value signals in the AGm. (**c**) Percentages of chosen value-coding neurons out of all, choice-, or outcome-coding neurons (*eq. 2*) were compared. There was no tendency for choice- or outcome-coding neurons to preferentially encode chosen value (χ^2-test, p=0.869 and 0.389, respectively; analysis based on the neural data during the first 1 s of the reward stage). This result suggests that chosen value signals were not merely due to elevated activity of those neurons related to reward consumption or movement toward a reward. (**d**) Population activity of chosen value-coding neurons associated with different block reward probabilities. Normalized population spike density functions (σ=50 ms) were constructed for those neurons that encoded chosen value during the first 1 s of the reward stage (n=69) according to the block reward probability of the goal chosen by the animal in each trial. Firing rates were overall similar unlike the neural activity previously reported for the neurons in the primate supplementary motor area and premotor cortex which was interpreted as motivation-related signals[1].



Supplementary Figure 7. Neural signals related to values that were computed with Rescorla-Wagner rule (simple or model-free RL). (**a**) Neural signals for decision value and chosen value. Same format as in Fig. 4a. (**b**) Neural signals for the animal's chosen action, reward or chosen value around the time of reward stage onset. Same format as in Fig. 5a. (**c**) Relationship between the standardized regression coefficients (SRC) for chosen value and reward for neural activity in the AGm during the first 1 s of the reward stage. Same format as in Supplementary Fig. 5.



Supplementary Figure 8. Isolation and classification of units. (**a**) An example of neural signals recorded with a tetrode. Each point in the scatter plot is a signal that exceeded the experimenter-defined threshold in at least one of 4 tetrode channels during 30 min of recording. The abscissa and ordinate indicate the energy of spike signals recorded through channels 1 and 2, respectively. Individual clusters are indicated in different colors. Four-channel average spike waveforms are shown for each cluster in corresponding color on the right. Calibration: 1 ms and 0.1 mV. (**b**) Unit classification. Recorded units were classified into broad-spiking (spike width ≥ 0.3 ms) and narrow-spiking (spike width < 0.3 ms) neurons based on the distribution of their spike width. Broad-spiking neurons (putative pyramidal cells) fired at low rates (AGm, 5.00±0.26; AGl, 5.29±0.53 Hz) and had high peak/valley ratios (AGm, 7.54±0.16; AGl, 7.69±0.36), whereas narrow-spiking neurons (putative interneurons) fired at high rates (AGm, 13.13±1.41; AGl, 14.58±2.29 Hz) and had low peak/valley ratios (AGm, 4.52±0.44; AGl, 3.94±0.88).

SUPPLEMENTARY TABLES

Supplementary Table 1. Results from statistical analyses for the lesion experiments. We performed two-way repeated measure ANOVA followed by paired *t*-tests to examine statistical significance of lesion effects on various behavioral measures. Group: lesion vs. sham (n=5 animals each). Phase: before vs. after lesions. Numbers indicate *p* values. Significant *p* values (< 0.05) are indicated in red. All analyses were based on behavioral data during 10 days before and after lesions except for *P(LS), 3 d* (behavioral data during the first 3 days were used as post-lesion data).

	ANOVA			***t*-test**	
	Group	**Phase**	**Interaction**	**Lesion (pre vs. post)**	**Sham (pre vs. post)**
Speed	0.790	0.225	0.642	0.408	0.145
Choice bias	0.016	0.190	0.069	0.121	0.490
P(LS)	0.040	0.634	0.007	0.073	0.038
P(LS), 3 d	0.004	0.059	0.004	0.019	0.219
P(high)	0.112	0.006	0.036	0.002	0.538
α	0.366	0.384	0.289	0.845	0.296
β	0.229	0.015	0.091	0.014	0.500
$P(Q_{low})$	0.121	0.070	0.038	0.038	0.783
$P(Q_{low}$, stay)	0.189	0.196	0.009	0.019	0.260
$P(Q_{low}$, switch)	0.682	0.222	0.779	0.483	0.337
-log(L)/N	0.028	0.013	0.060	0.009	0.586
Prediction	0.045	0.038	0.037	0.003	0.992

Supplementary Table 2. Model comparison. Performances of the SP model[2] and Rescorla-Wagner rule[3], which is a model-free RL algorithm that updates only the value of chosen action according to a choice outcome, were compared. The accuracy to predict the animal's actual choice (% correct), Akaike's information criteria (AIC) and Bayesian information criteria (BIC) are shown (mean±SD, *n*=3 animals). Model prediction of the animal's choice was assessed by applying a leave-one-out cross-validation procedure to the behavioral data across different sessions.

	AIC	BIC	Prediction (% correct)
Rescorla-Wagner rule	2,628±118	2,639±118	64.6±2.0
SP model	2,476±58	2,493±58	68.1±1.3

SUPPLEMENTARY NOTE

Stacked Probability Algorithm

Detailed procedures for calculating action values are described in our previous report[2]. Briefly, Rescorla-Wagner rule was modified to estimate arming probability (referred to as 'stacked' arming probability) that varies with run length (the number of consecutive alternative choices). The stacked arming probability associated with an action (i.e., left or right goal choice) in the *i*-th trial [$S^{est}_{action}(i)$] was calculated by combining estimated block arming probability [$A^{est}_{action}(i)$] and estimated stack parameter [$X^{est}_{action}(i)$], which increases with the number of consecutive alternative choices, as follows:

$$S^{est}_{action}(i) = A^{est}_{action}(i) X^{est}_{action}(i),$$

where $X^{est}_{L}(i) = \sum_{n=0}^{n_R(i)} \left(1 - A^{est}_{L}\right)^n$ if $action = L$

$X^{est}_{R}(i) = \sum_{n=0}^{n_L(i)} \left(1 - A^{est}_{R}\right)^n$ else.

The estimated block arming probability was updated according to the Rescorla-Wagner rule as follows:

$$A^{est}_{action}(i+1) = (1-\alpha) A^{est}_{action}(i) + \frac{\alpha}{X^{est}_{action}(i)} \cdot r(i) \quad \text{if } action = action(i)$$

$$A^{est}_{action}(i+1) = A^{est}_{action}(i) \quad \text{else.}$$

After the block arming probability was estimated, the stack parameter was updated as follows:

$$X^{est}_{action}(i+1) = 1 \quad \text{if } action = action(i)$$

$$X^{est}_{action}(i+1) = 1 + X^{est}_{action}(i) \cdot \left(1 - A^{est}_{action}(i)\right) \quad \text{else.}$$

The choice of action was determined by the softmax action selection rule as the following:

$$P_L(i) = \frac{1}{1 + \exp(-\beta R_D(i))},$$

where $P_L(i)$ is the probability for selecting the left goal, β is the inverse temperature that defines the degree of exploration in action selection, and $R_D(i)$ is the estimated difference in local arming probabilities which was represented as

$$R_D(i) = A^{est}_{L}(i) X^{est}_{L}(i) - A^{est}_{R}(i) X^{est}_{R}(i).$$

REFERENCES

1. Roesch, M.R. & Olson, C.R. Impact of expected reward on neuronal activity in prefrontal cortex, frontal and supplementary eye fields and premotor cortex. *J. Neurophysiol.* **90**, 1766-1789 (2003).

2. Huh, N., Jo, S., Kim, H., Sul, J.H. & Jung, M.W. Model-based reinforcement learning under concurrent schedules of reinforcement in rodents. *Learn. Mem.* **16**, 315-323 (2009).

3. Rescorla, R. & Wagner, A. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. in *Classical Conditioning II: Current Research and Theory* (eds. Black, A. & Prokasy, W.) 64-99 (Appleton, New York, 1972).