nature portfolio

https://doi.org/10.1038/s41586-025-09466-1

Supplementary information

Rate and noise in human amygdala drive increased exploration in aversive learning

In the format provided by the authors and unedited

Rate and noise in human amygdala drive increased exploration in aversive learning

Tamar Reitich-Stolero[1], Kristoffer C. Aberg[1,2], Dean Halperin[1], Carmel Ariel[1], Genela Morris[6], Lilach Goldstein[3,4], Firas Fahoum[3,4], Ido Strauss [5,6,7], Rony Paz[1,2,7]

[1] Department of Brain Sciences, Weizmann Institute of Science, Rehovot 7610001, Israel

[2] Azrieli Institute for Brain and Neural Sciences, Weizmann Institute of Science, Rehovot 7610001, Israel

[3] Epilepsy Unit, Neurological Institute, Tel Aviv Medical Center, Tel Aviv 6423906, Israel

[4] Faculty of Medicine, Tel Aviv University, Tel Aviv 6997801, Israel

[5] Functional Neurosurgery Unit, Tel Aviv Sourasky Medical Center, Tel Aviv 6423906, Israel

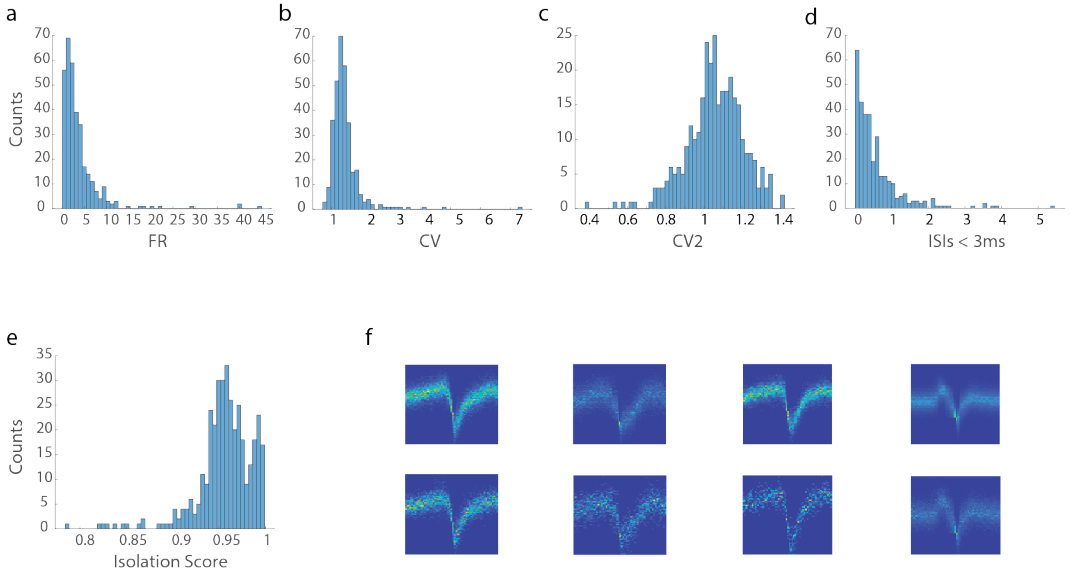
[6] Department of Neurology and Neurosurgery, School of Medicine, Tel Aviv University, Tel Aviv 6997801, Israel

[7] These authors contributed equally

Correspondence: Rony Paz (rony.paz@weizmann.ac.il)

Supplementary figures and tables

Supplementary Fig. 1. Spike-sorting quality



(a) Histogram of firing rates (FR) across all recorded neurons, mean FR = 3.8.

(b-c) Histogram of coefficient of variance of the ISI distribution (CV_{ISI}) and the modified coefficient of variance ($CV_{ISI}2$). The mean score is significantly larger than 1 in both CV_{ISI} and $CV_{ISI}2$, suggesting irregular firing patterns (CV_{ISI}: $t_{368} = 14.45, p < 0.001, d = 0.81$; $CV_{ISI}2$: , $t_{319} = -6.7, p < 0.001, d = 0.37$).

(d) Histogram of the percentage of ISIs that are shorter than 3ms in each of the recorded units. In the majority of units (58%) short ISI proportion is less than 0.5%.

(e) Histogram of isolation scores (mean = 0.96). See methods for definition and references.

(f) Spike density plots of the four example single neurons (See Fig.3a, columns) recorded in the two main conditions of the experiment (upper row: exploitation trials; bottom row: exploration trials).

Supplementary Table 1. Single unit recording locations by participants and session

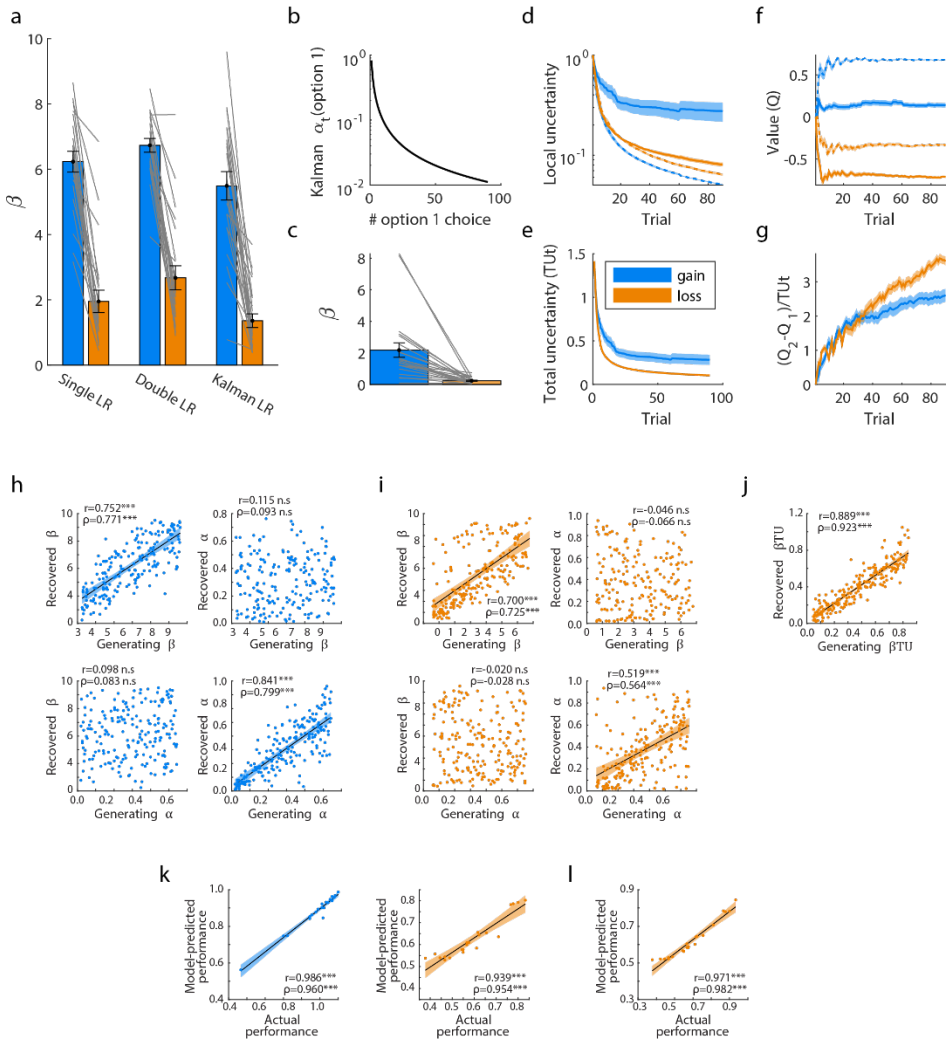
Session number, participant number, sex and the number of recorded units from each electrode-location.

# Session		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
# Participant		1	2	2	3	4	5	6	7	8	8	9	9	10	11	12	12	13	14	15	16	16	17
Sex (M/F)		M	M	M	F	M	M	M	M	M	M	M	M	M	M	F	F	M	M	M	F	F	M
Temporal cortex (78)	TP	0	5	2	0	0	0	0	0	4	3	0	0	12	15	0	0	7	0	4	0	0	0
	Insula	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	PC	0	0	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0
	MC	3	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
Amygdala (44)	Amy	0	0	0	1	0	0	0	1	0	0	2	4	9	4	0	0	6	0	13	2	0	2
Sensory motor cortices (63)	STG	0	0	0	0	0	0	4	0	0	0	0	0	0	4	0	0	3	0	13	0	0	0
	postT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	3	0
	MTG	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O	0	0	0	0	0	0	0	0	5	2	0	0	0	0	0	0	0	0	0	0	0	0
	PO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
	TO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	2	0
	SMA	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Frontal cortices (109)	AF	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	OF	3	0	0	0	0	0	8	0	0	0	5	0	0	8	1	0	0	2	3	6	0	0
	MF	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ASF	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	SF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	0	0	0	0	0	0
	OFAC	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	SFG	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0
	FPolar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0
	F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
	AC	0	0	0	0	0	1	0	2	0	0	0	0	0	0	7	0	0	0	0	5	7	0
Hippocampal regions (88)	AH	0	0	0	0	0	0	0	0	0	0	7	7	0	0	0	0	0	0	0	5	6	0
	MH	0	0	1	0	0	0	0	2	0	0	1	1	1	0	0	0	0	0	0	0	0	0
	EC	0	0	0	0	0	0	0	0	4	2	0	0	0	0	0	0	0	0	13	0	0	0
	HC	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3	0	0	2	0	12	4	0
	PHG	0	0	0	0	0	0	7	0	4	0	0	0	5	0	0	0	0	0	0	0	0	0

Supplementary Table 2. List of abbreviations

AC	anterior cingulate	HC	hippocampus	OF	orbitofrontal	SF	superior frontal	TP	temporal pole
AF	anterior frontal	MC	medial cingulate	OFAC	orbitofrontal anterior cingulate	SMA	supplementary motor area		
AH	anterior hippocampus	MF	medial frontal	PC	posterior cingulate	SFG	sup. frontal gyrus	ant	anterior
ASF	ant. superior frontal	MH	medial hippocampus	PHG	para hippocampal gyrus	STG	sup. temporal gyrus	mid	medial
EC	entorhinal cortex	MTG	medial temporal gyrus	PO	parietal occipital	T	temporal	post	posterior
F	Frontal/ frontal cortex	O	occipital lobe	pT	posterior temporal	TO	temporal occipital	sup	superior

Supplementary Fig. 2. Reinforcement learning models indicate higher sensitivity to uncertainty in loss



(a) Mean (bars) and single participant values (gray lines) of the inverse-exploration term (inverse soft-max temperature, β_Q) for a Q-learning model with a single learning-rate (LR), two LRs, and Kalman filter LR. The inverse-exploration term in all models was smaller for the loss choices compared to gain choices (n=22, double LR: $Z = 3.74, p < 10^{-3}$, Kalman filter LR: $Z = 4.01, p < 10^{-4}$; see main text for single LR).

(b-g) The model that describes best the behavior during loss has sensitivity to total uncertainty (Q/TU) with a Kalman Filter LR (Fig. 2k)**.**

(b) The LR for a single option as a function of the number of choices of that option (the LR depends only on the number of choices).

(c) Same as in (a) for this model ($\beta_{Q/TU}$).

(d) Local uncertainty of each option as a function of trial, averaged across all participants. Because it was sampled less, the inferior option (full lines) has higher local uncertainty than the superior option (dashed lines).

(e) Total uncertainty in the gain and loss conditions is the mean of the local uncertainties for both options, averaged across all participants, as a function of trial. The total uncertainty of the loss condition tends to be lower as both options were sampled frequently.

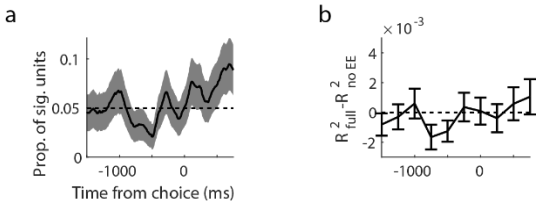
(f) Average value (Q) across participants as a function of trial for each option in loss and gain.

(g) The average difference in the values (Q) of the two options scaled by the total uncertainty, for loss and gain.

(h, i, j) Parameter recovery. Shown are correlations between the generating and the recovered parameters for the 'Single LR-Q' model in the Gain condition **(h)**; for the 'Single LR-Q' model in the Loss condition **(i)**; and correlation between the generating and recovered β_{TU} parameter for the 'Kalman-filter LR-Q/TU' model in the Loss condition **(j)**. Significant correlations are obtained only between generating and recovered parameters of the same type (e.g. between generating and recovered β in the Gain condition; top left panel in h), but not between different parameters (e.g. between generating β and recovered α in the Gain condition; top right panel in a). These results, showing little shared variance between model-parameters, suggest that each parameter can be meaningfully interpreted as governing specific aspects of behavior. n=200 simulations in all; ***p<0.001.

(k, l) Actual versus model-predicted inter-individual performances (n=17). Shown is the correlation between actual and model-fitted (Single LR-Q model) performance in the Gain condition (**k**, left) and in the Loss condition (**k**, right); and the correlation between actual and model-fitted (Kalman-filter LR-Q/TU model) performance in the Loss condition **(l**). ***p<0.001.

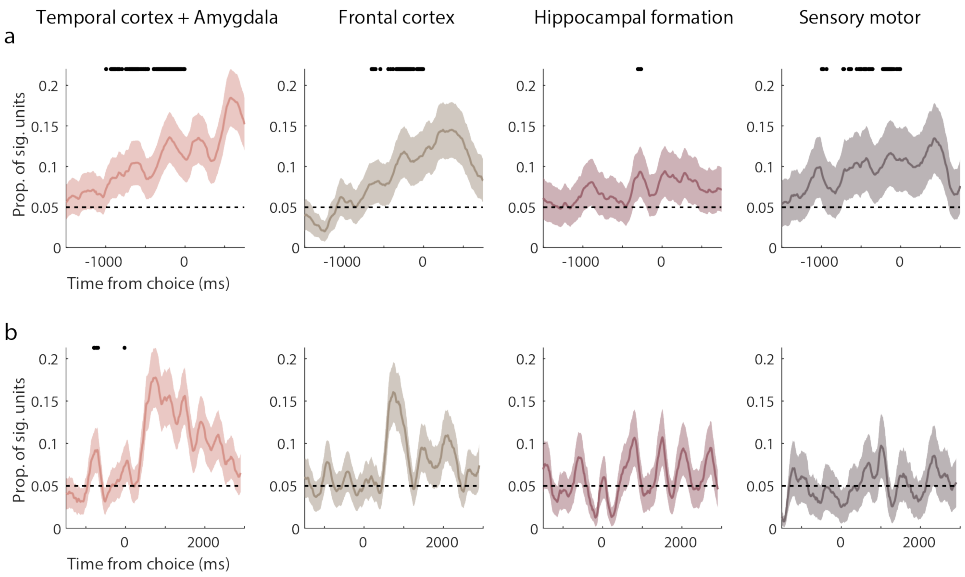
Supplementary Fig. 3. Correct-Incorrect does not account for the exploration-exploitation signal



(a) The proportion of units in the temporal cortex and amygdala with a significant effect for correct-incorrect trials was at chance level (same model as in Fig. 3e, ANOVA, accounting for correct-incorrect, gain-loss and the sign of the prediction error).

(b) The difference between explained variability (R^2) in models with and without the correct-incorrect predictor.

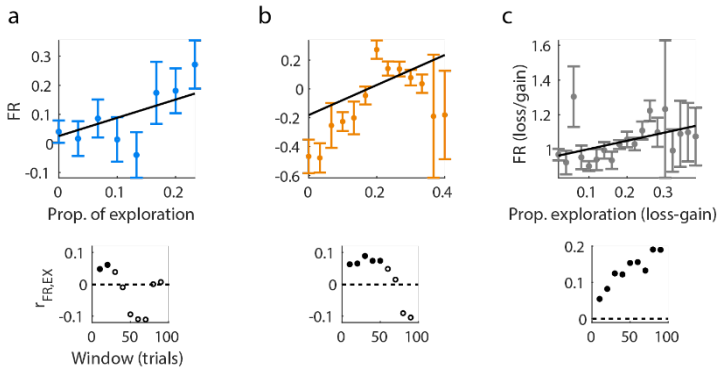
Supplementary Fig. 4. Responses to choice identity and the sign of the prediction-error



(a) Proportion of units with a significant effect for choice identity (as in Fig. 3e, ANOVA model, accounting for EE, choice identity and the sign of the prediction error, FDR corrected).

(b) Proportion of units with a significance effect for the sign of the prediction error. As expected, PE signal appears mainly after the choice/feedback.

Supplementary Fig. 5. Activity of single neurons correlates with the probability of exploration.

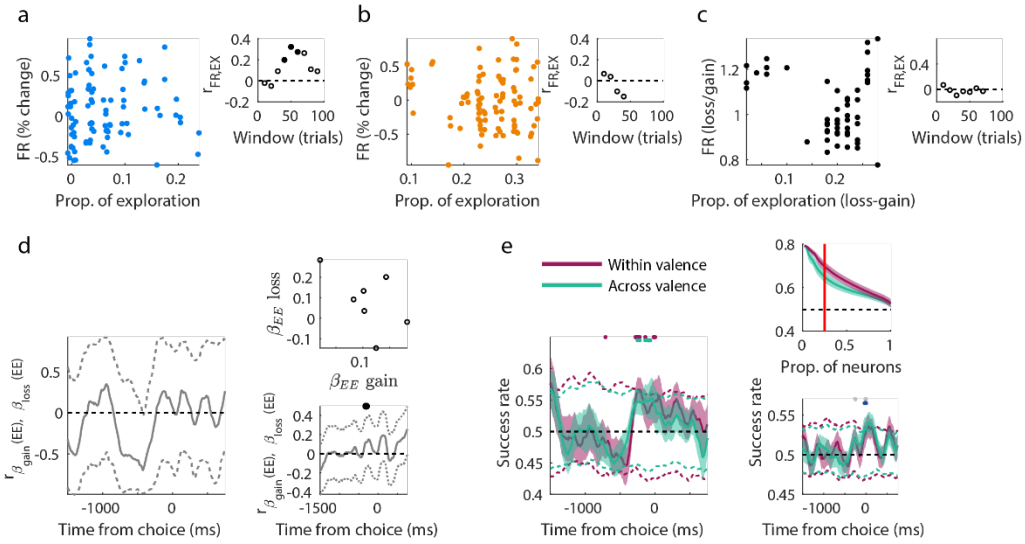


(a) Activity is correlated with the proportion of behavioral exploration in gain trials. The correlation is computed over all single-neurons in the amygdala and temporal cortex in a 30-trial window (namely, each data point is the activity of one neuron averaged over 30 trials and the proportion of exploration in these trials). For presentation purposes only, the activity is averaged in bins of proportion of exploration (bins are determined by the window size). Inset shows the consistency of the positive correlation for different window sizes (full circles: $p<0.05$).

(b) as in (a) for loss trials.

(c) as in (a,b) for the difference between loss and gain.

Supplementary Fig. 6. Sensory-motor patterns do not account for the similarities across gain and loss in exploration signals



Panels as in Fig.4 for sensory-motor neurons.

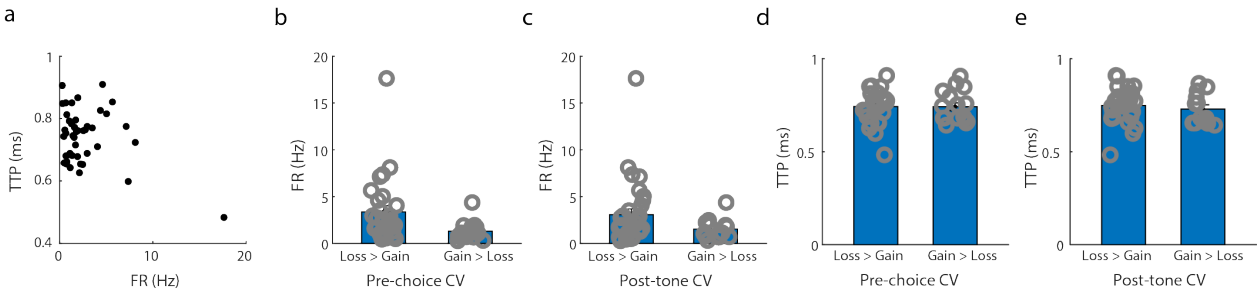
(a-b) FR shows some significant increase with the proportion of behavioral exploration (across 30 trials) in gain trials (a) but not in loss trials (b).

(c) FR ratios between loss and gain is not correlated with the behavioral difference.

(d) There is no positive correlation between $\beta_{\text{gain}}(EE)$ and $\beta_{\text{loss}}(EE)$. Confidence intervals in dashed lines ($\alpha = 0.05$ for a one-tailed significance test). Right top: the relationship shown for pre-choice activity (averaged over 400ms prior to choice; only neurons with significant response to exploration-exploitation). Right bottom: for all neurons with sufficient sampling in the sensory-motor cortex ($n = 40$).

(e) Decoding accuracy of exploration-exploitation within valence (maroon) and across valence (mint green). Shown for the 25% top performance neurons (main) and for all neurons (right bottom). Right top: average decoding accuracy as a function of the proportion of neurons taken, adding neurons from the best to worst decoding performance.

Supplementary Fig. 7. Principal cells and putative interneurons contribute similarly.

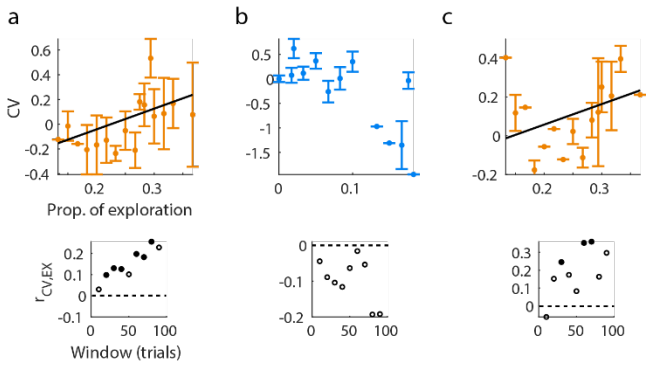


(a) Trough-to-peak (TTP) time interval as a function of ongoing FR (measured throughout the experiment, independent of behavioral events) of all neurons recorded from the amygdala. There were no clear clusters of putative interneurons and pyramidal cells, and it seems that most cells are pyramidal.

(b-c) Overall FR (mean and SE) of single amygdala neurons that showed larger loss CV compared to gain CV the pre-choice period (b) and after tone presentation (c). There was a significantly higher FR for neurons with higher post-tone CV in loss (ranksum test, $n = 44, ranksum = 740, z = 2.67, p < 0.05$).

(d-e) Same as b-c, comparing TTP between the groups. No comparison was significant (all p>0.05).

Supplementary Fig. 8. Noise (CV) across single neurons increases with the probability of exploration.

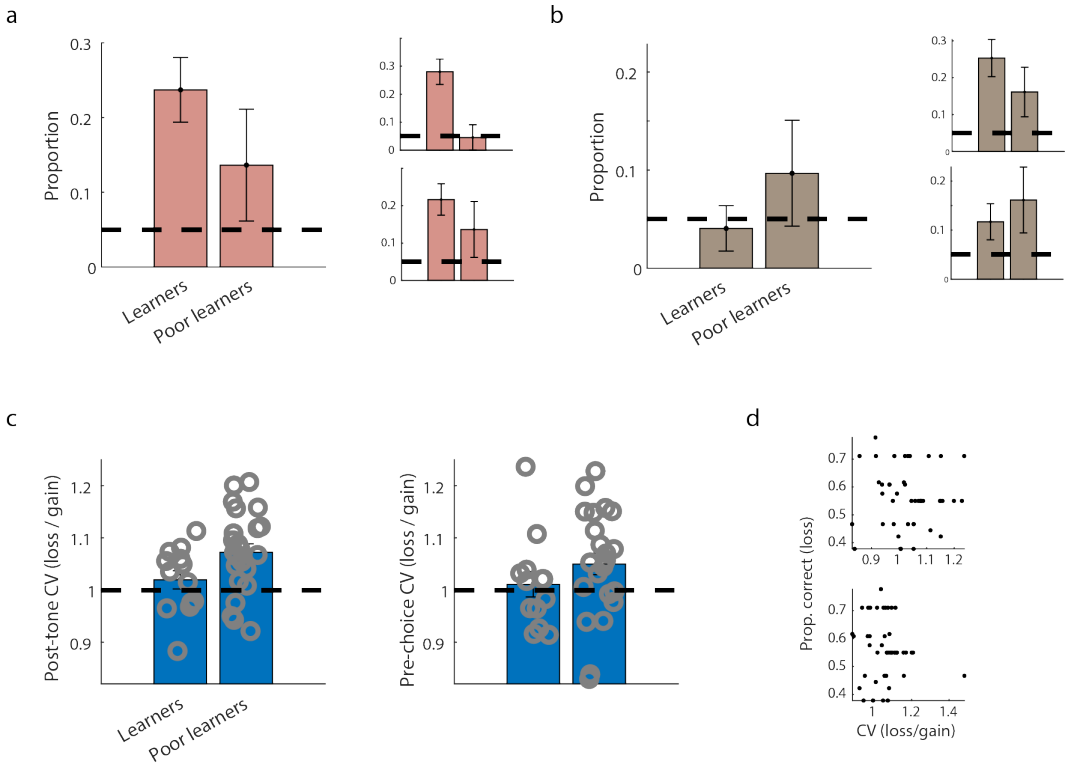


(a) Noise (CV) increases with the proportion of behavioral exploration in loss trials, averaged over all amygdala neurons (same format as in Supp. Fig.5, $n = 152$, $r_{sp} = 0.19, p = 0.008$). Inset shows the consistency and robustness of the correlation for different window sizes (filled circles: p<0.05).

(b) Noise (CV) does not correlate with the proportion of exploration in gain trials ($n = 152, r_{sp} = -0.016, p > 0.5$).

(c) Same as (a) but calculated only on exploitation loss trials**.**

Supplementary Fig. 9. FR and noise signals are robust across significant learners and poor learners



(a) Similar proportion of units in the amygdala and temporal cortex with a significant EE response ([-500,0] pre-choice) in participants with significant learning performance and with non-significant learning ($n = 97,22$ respectively). Insets: same for choice identity (top, [-500, 0] pre-choice) and for PE (bottom, [0,500] post-choice). Significant units were tested with analysis of variance (ANOVA model, as in Fig.3e, binomial criterion for significance across time bins, $p_0 = 0.05, \alpha = 0.05$, Error bars mark SE). There was no significant difference in the proportion of units signaling EE and PE, while there was an expected higher proportion of choice-identity units in significant learners.

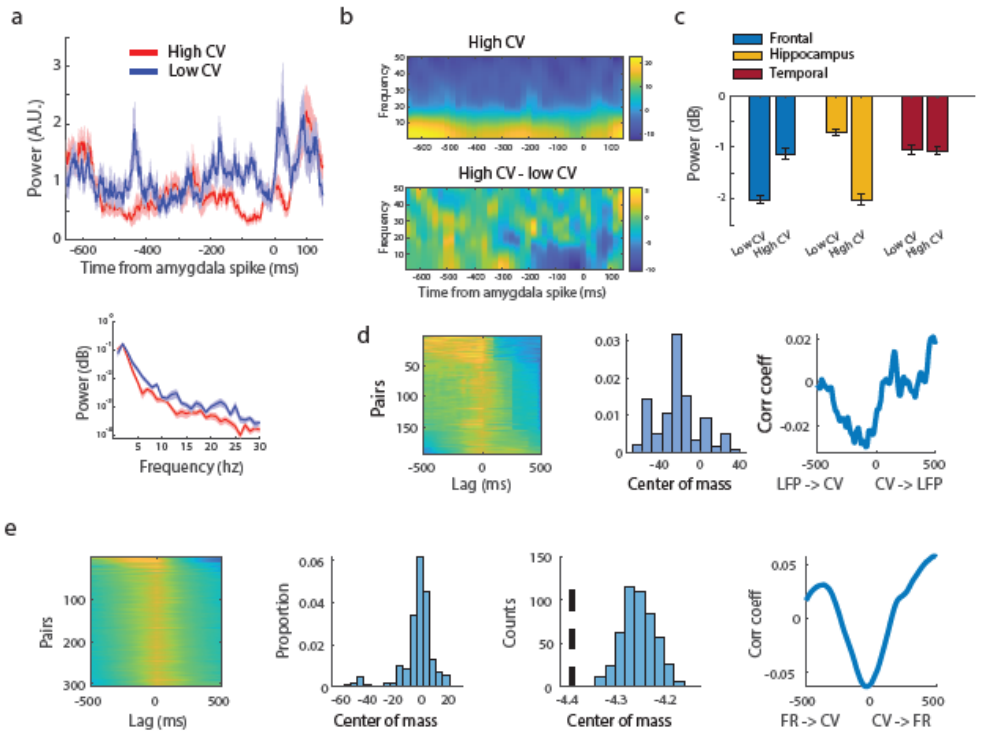
(b) Same as A for the frontal cortex units (n=74,31 for significant and non-significant learners), confirming the existence of learning signals (choice identity, PE) in both groups of participants, but no EE signal in both groups.

(c) CV ratio (loss/gain) after tone presentation (left, [0,2000] post-tone) and pre-choice (right, [-1000,0] pre-choice). The CV was higher in loss than in gain in all groups and all conditions, with no significant difference across significant and non-significant learners ($n = 13,24$, respectively, rank sum test, post tone: $ranksum = 483, z = 1.88, p = 0.06$, pre choice:, $ranksum = 511, z = 1.73, p = 0.08$).

(d) There was no correlation between proportion of correct responses and CV ratio (loss/gain), both in post-tone activity (top, $r = -0.11, p = 0.48$) and in pre-choice activity (bottom, $r = -0.14, p = 0.36$). Hence, noise levels

are not directly related to high/poor performance in the loss condition, while they are directly related to exploration. This finding reinforces the idea that the lower performance results from higher exploration.

Supplementary Fig. 10. Lower hippocampal activity precedes higher amygdala noise



(a) Hippocampal power was lower just before an amygdala spike in high CV trials compared to low CV trials. Shown is spike-triggered-average LFP, separately for amygdala spikes from high CV and from low CV loss trials. Top: overall power around amygdala spikes. Bottom: power spectral density during 500ms preceding amygdala spike.

(b) LFP spectrograms as a function of time from amygdala spike in loss trials. Top: during high CV trials. Bottom: the difference between high and low CV trials (high-low).

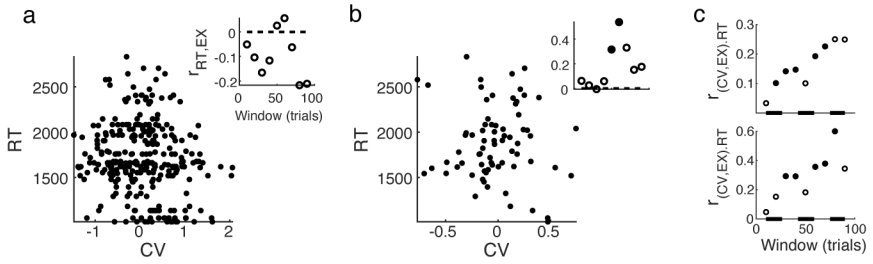
(c) LFP power averaged over the 500ms preceding amygdala spikes, separately for high CV and low CV loss trials. Only the hippocampus showed a significant difference (two-tailed paired t-test: $n = 36, t_{35} = 2.15, \mathrm{p} < 0.05, d = 0.48$).

(d) Lower LFP power is associated with and precedes an increase in amygdala noise. Shown are cross-correlations between pairs of CV of amygdala neurons and hippocampal LFP power. Left: cross-correlation strength as a function of lag (ms) for each pair of simultaneously recorded neuron-LFP. Middle: center of mass for all pairs (one sample t-test: $n = 195, t = 14.8, p < 0.001, d_{cohen} = 1.05$). Right: Normalized cross-correlation (Pearson

correlation coefficient) averaged over all pairs ($n = 195$). CV in amygdala single-neurons and LFP in the hippocampus were taken during two seconds of pre-choice activity.

(e) Lower hippocampal FR is associated with and precedes an increase in amygdala noise. Shown are cross-correlations between the CV of amygdala neurons and the firing rate (FR) of single hippocampal neurons. Left: cross-correlation strength as a function of lag (ms) for each pair of simultaneously recorded neurons. Middle Left: center of mass for all pairs (one-sample t-test: $n = 314, t = 8.6, p < 0.001, d_{cohen} = 0.49$). Middle Right: The mean center of mass (black dashed line) is significantly smaller than shuffled data (distribution computed from shuffled data sets preserving activity within trials, Monte Carlo p-value, $p < 0.005$). Right: Normalized cross-correlation (Pearson correlation coefficient) between amygdala CV and hippocampal FR, averaged over all pairs ($n = 314$).

Supplementary Fig. 11. Reaction times do not account for higher CV and exploration in loss

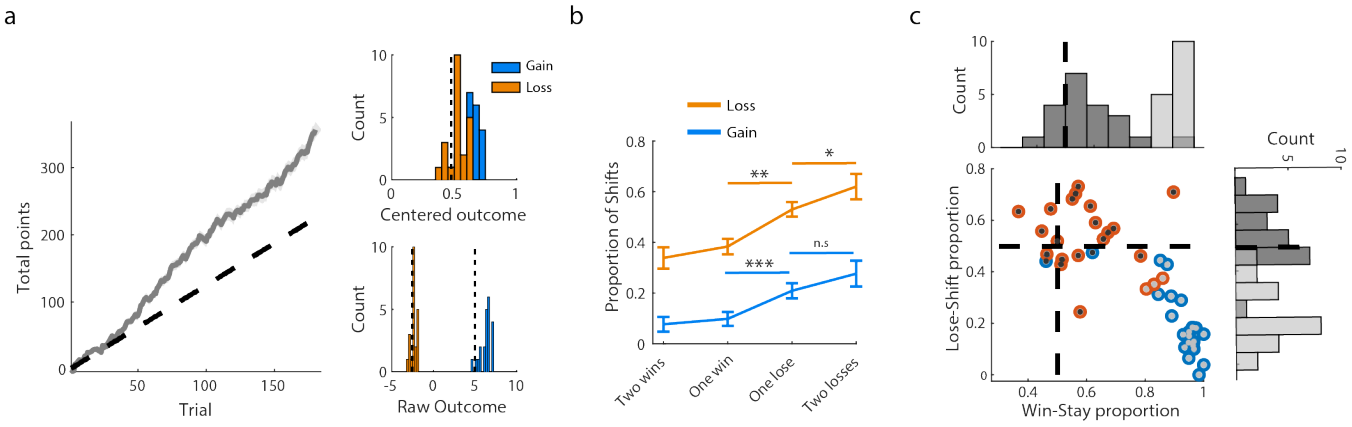


(a) There was no significant correlation between noise (CV of single neurons) and RT in loss trials ($n = 294, r_{sp} = -0.12, p > 0.5$; Inset shows consistency across window size).

(b) There was no significant correlation between noise (CV, averaged over all neurons recorded simultaneously) and RT in loss trials ($n = 72, r_{sp} = 0.06, p > 0.05$). Inset is as in a.

(c) The correlation between CV and probability of exploration remains significant for most window sizes after controlling for the RT (partial correlation; full circles are significant at p<0.05). Top: CV of single-neurons; bottom: CV averaged over all neurons recorded simultaneously.

Supplementary Fig. 12. Outcome and feedback sensitivity do not account for higher exploration in loss



(a) Total points of participants as a function of trials (Mean and SE). The total points score was shown at every feedback, with a synthetic high-score table presented at the start to increase motivation. Insets show average outcome for the last 30 gain and loss trials, with both higher than chance level (dashed line). Top: Centered outcome (using min-max scaling to enable comparison). Bottom: Raw scores. Due to the higher exploration and lower proportion of correct choices in loss trials, average outcomes were lower.

(b) Win-Stay-Lose-Shift (WSLS) behavior: Participants showed increased switching following fewer wins and more losses (comparing 2/1 consecutive wins, 1/2 consecutive loss, 2-way repeated measure ANOVA, main effect for previous feedback [n=22 within each group]: $F(13,63) = 31.68, p < 10^{-11}, \eta^2 = 0.06$), and a higher overall switching rate in loss compared to gain trials (main effect of valence, $F(1,21) = 77.9, p < 10^{-7}, \eta^2 = 0.14$), with no significant interaction ($F(3,63) = 0.65, p > 0.5, \eta^2 = 0.001$). This indicates that participants were sensitive to prior feedback in both gain and loss contexts, with a greater overall tendency to switch in loss context. Multiple comparison analysis showed significantly higher probability of shifting after losses compared to wins in both contexts (one tailed Tukey-Kramer multiple comparisons test: gain: $p < 10^{-4}$, loss: $p < 0.01$) and higher probability of shifting after 2 losses compared to 1 loss in the loss condition ($p < 0.05$).

(c) Proportion of WSLS behavior for individual participants in gain (blue) and (orange) trials. Dashed line mark random behavior (p=0.5). K-means clustering (k=2) revealed two groups with putative strategies. One (light gray, n=23) with high win-stay (right-tailed sign-rank test for $p_0 = 0.5, median = 0.95, Z = 4.18, p < 10^{-4}$) and low lose-shift behavior ($median = 0.16, Z = -4.2, p > 0.5$), that included almost all gain blocks and only 3 loss blocks. The other strategy (dark gray, n=21) showed more moderate (yet higher than chance) win-stay behavior ($median = 0.57, Z = 2.63, p < 0.01$) and moderate (higher than chance) lose-shift behavior ($median =$

$0.55, Z = 1.66, p < 0.05$). In both groups, lose-shift proportion was higher than win-shift proportion (group 1: $Z = 4.12\ p < 10^{-4}$; group 2: $Z = 2.61, p < 0.01$), again supporting sensitivity of choices to feedback. Insets show marginal histograms.