Supplementary Note 1: Limitation of the multiplicative strategy in differential weighting of reward information.

In a model based on the multiplicative strategy, the information about reward probability and magnitudes is combined to construct subjective reward value. Although normative, this strategy for combination of reward information limits independent adjustments of the influence of reward probability and magnitude on reward value and the ensuing choice. This is because based on this strategy, the only way that the relative weighting of reward probability and magnitude can be adjusted is through changing the utility and/or probability weighting functions, both which are commonly assumed to be fixed for a given set of options.

Nevertheless, an exponential transformation can turn multiplication into summation ($\exp(a) \times \exp(b) = \exp(a + b)$). Therefore, if one assumes an exponential form for the representations of reward attributes (i.e., if objective reward attributes can be encoded exponentially by neural activity and such signals can be easily multiplied), a multiplicative model becomes an additive one. Consequently, even in a model based on a multiplicative strategy, reward probability and magnitude can differentially influence subjective value and subsequent choice if their exponential transformations are adjusted.

Although this scenario seems plausible, the combination of reward information for evaluation cannot be considered separately from the subsequent decision-making processes [1]. Importantly, an additive strategy implies decision making based on direct comparisons of reward attributes (probability and magnitude). In contrast, if one assumes evaluation based on a multiplicative strategy, subsequent decision making would not be based on direct comparisons of reward attributes of alternative options. Therefore, the fundamental difference between the additive and multiplicative strategies is whether different reward attributes of each option are fused before the onset of decision-making processes. A multiplicative model inherently requires such fusion whereas an additive model can also accommodate decision making based on direct comparisons of attributes in each dimension separately [2].

[1] Stewart, N. Information integration in risky choice: Identification and stability. *Front. Psychol.* **2**, 301 (2011).

[2] Tversky, A. Intransitivity of preferences. *Psychol. Rev.* **76**, 31 (1969).

Supplementary Note 2: **Direct behavioral evidence for adjustments of learning and valuation strategy to volatility.**

To measure behavioral adjustments to volatility of the environment directly (and not based on the fit of choice data), we calculated the log odds of choosing the better (option with the higher probability of reward) vs. the worse option. Assuming that the subjective value of each gamble is an additive function of its estimated reward probability and reward magnitude, and that the probability of selection between the two options is a sigmoid function of the difference in their subjective values, this log odds is equal to:

$$\log odds\left(p_{B(m_B,m_W)}\right) = log\left(\frac{p_{B(m_B,m_W)}}{1-p_{B(m_B,m_W)}}\right) = \beta_m(m_B - m_W) + \beta_p(p_B - p_W) \quad \text{(Eq. S1)}$$

where β_p and β_m indicate the relative weights of reward probability and magnitude on subjective value, $p_{B(m_B,m_W)}$ denotes the probability of choosing the better option, and m_B and m_W are the reward magnitudes for the better and worse options, respectively,. The better option could be assigned with the smaller or larger reward magnitude, but reward magnitudes could also be equal for the two options. This results in three types of trials for which $\log odds\left(p_{B(m_B,m_W)}\right)$ can be computed: $\log odds\ (p_{B(m,M)})$, $\log odds\ (p_{B(M,m)})$, and $\log odds\ (p_{B(m',m')})$, where *M, m*, and *m'* denote the larger, smaller, and the same reward outcomes of the two options. These three quantities can be used to estimate the overall impact of reward probability, $\tilde{b}_p$, and the weight of reward magnitude on choice, b_m, as follows:

$$\tilde{b}_p = \frac{1}{n_u + n_e}\left(\sum_{\{gamble\,\in\,Su\}}\left(\frac{\log odds\left(p_{B(m,M)}\right) + \log odds\left(p_{B(M,m)}\right)}{2}\right) + \sum_{\{gamble\,\in\,Se\}} \log odds\left(p_{B(m',m')}\right)\right)$$

$$= \beta_p(p_B - p_W) \quad \text{(Eq. S2)}$$

$$b_m = \frac{1}{n_u}\left(\Sigma_{\{gamble\,\in\,Su\}}\left(\log odds\left(p_{B(m,M)}\right) - \log odds\left(p_{B(M,m)}\right)\right)\right) / \left(2(m - M)\right) = \beta_m \quad \text{(Eq. S3)}$$

where Su denotes a subset of option pairs with unequal reward magnitudes ($Su = \{(1,2), (1,4), (2,4), (1,8)\}$), Se denotes a subset of pairs with equal reward magnitudes ($Se = \{(1,1), (4,4)\}$), and n_u and n_e are the total number of pairs in Su and Se, resepively. To obtain smoother estimates for $\tilde{b}_p$ and b_m, we calculated the probabilities of choosing the better option by combining data from all sessions and using a running average over time with the size of 4 trials.

Importantly, the ratio of the overall impact of reward probability to the weight of reward magnitude, $\frac{\tilde{b}_p}{b_m}$ ($= \frac{\beta_p(p_B - p_W)}{\beta_m} = \frac{(p_B - p_W)}{\beta_m/\beta_p}$), reflects both learning of reward probabilities ($p_B - p_W$) and the relative weighting of reward information (β_m/β_p), which we refer to as magnitude-to-probability weighting. Fitting choice behavior using a time-dependent additive model (see below for details) did not provide any evidence that β_m/β_p changed over time during each reversal (Supplementary Fig. 6b). Therefore, temporal dynamics of $\frac{\tilde{b}_p}{b_m}$ could capture the dynamics of learning whereas the difference between the steady state and initial value could assess the relative weighting of reward magnitude to probability.

Therefore, we next fit the overall impact of reward probability to the weight of reward magnitude ($\frac{\tilde{b}_p}{b_m}$) using a rising exponential function as follows:

$$y = y_\infty + (y_0 - y_\infty)e^{-t/\tau} \qquad \text{(Eq. S4)}$$

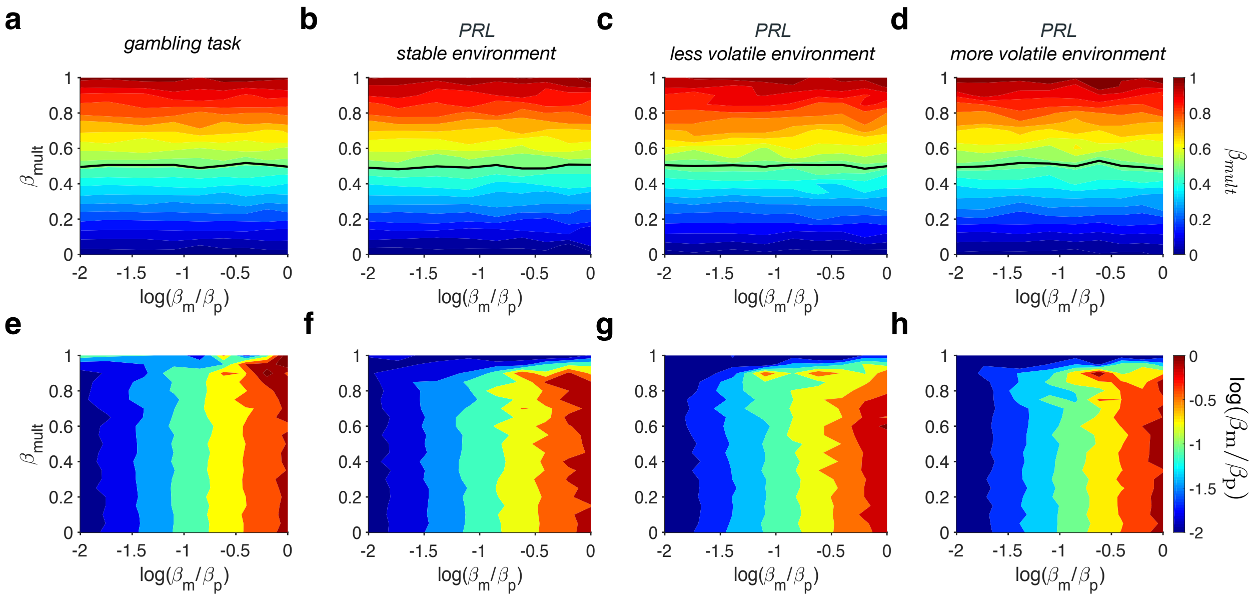
where y_∞ is the steady state, y_0 is the initial value and τ is the time constant. We found that $\frac{\tilde{b}_p}{b_m}$ reaches 95% of its steady state value at 10 ($= 3\tau$) trials in the more volatile environment in comparison with 15 ($= 3\tau$) trials in the less volatile environment (Supplementary Fig. 6c).

These results demonstrate that learning was slow and depended on volatility of the environment. Moreover, we found that the difference between the steady state and initial value was smaller for the more volatile compared to the less volatile environment ($y_\infty - y_0 = 5.9$ and 5.6 in the less and more volatile environments, respectively), dovetailing our results on the changes in relative in magnitude-to-probability weighting. Together, these results provide direct evidence for adjustments of learning and choice behavior to volatility of the environment.

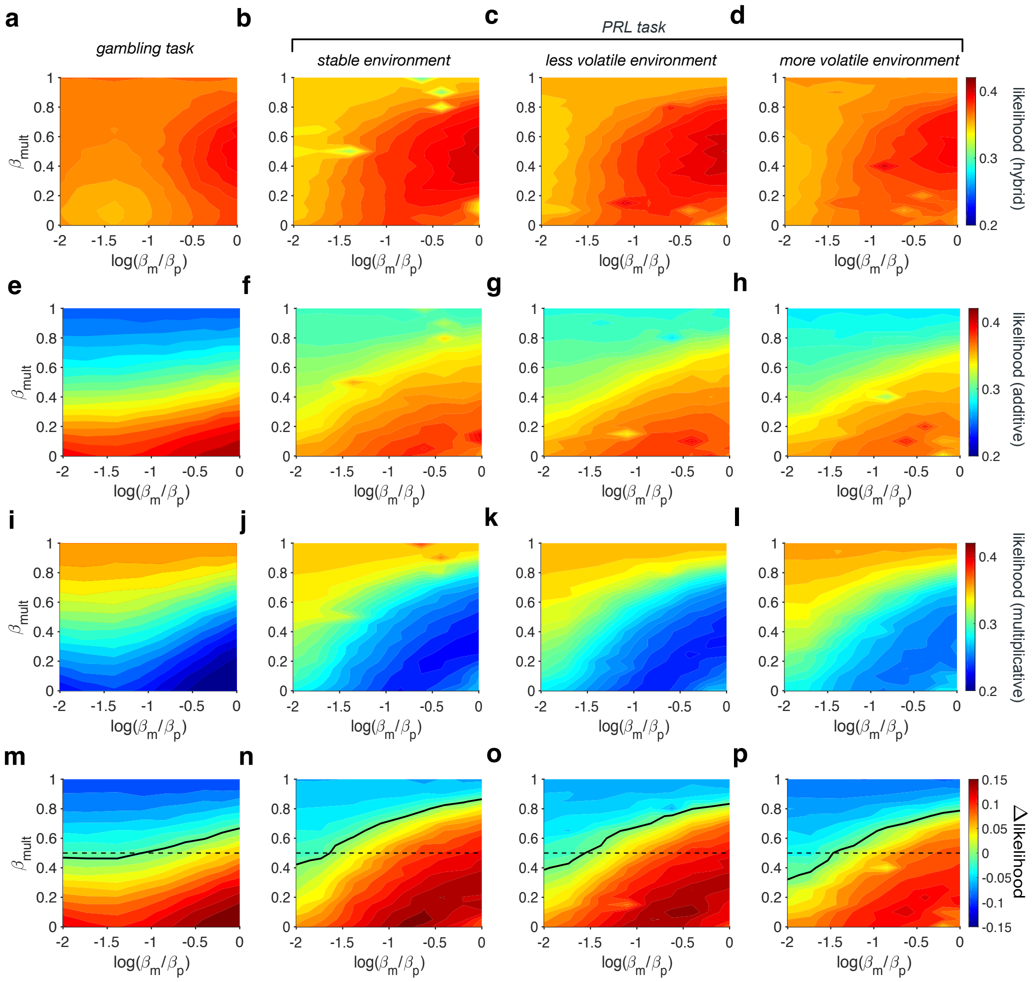
Fitting with time-dependent additive model. To test whether the relative weighting of reward information (β_m/β_p) changes over time, we fit choice behavior using a time-dependent additive model (t-additive). In this model, we allowed different weights for reward probability in the first and second half of trials in each reversal as follows:

$$SV_L = \begin{cases} \beta_m u(m_L) + \beta_{p(first)} w(p_L), & if\ t < L/2 \\ \beta_m u(m_L) + \beta_{p(second)} w(p_L), & if\ t > L/2 \end{cases} \qquad \text{(Eq. S5)}$$

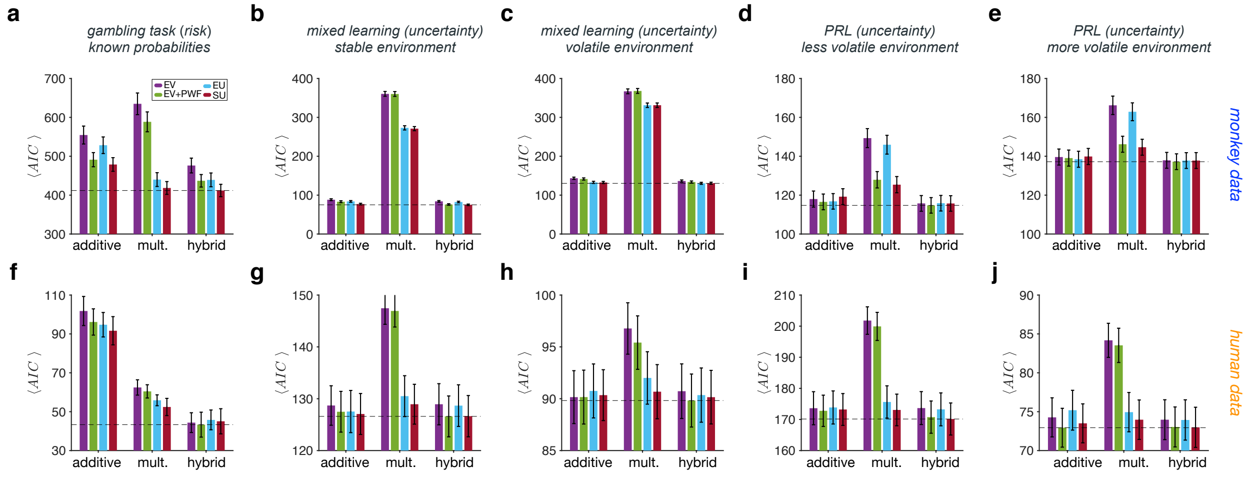
where t represents the trial number within a session and L is the block length. We then compared the results of fitting choice behavior using the time-dependent vs. simple additive models. Overall, we found that time-dependent weighting did not improve the fit as the simple additive model provided a better fit in both environments (Supplementary Fig. 6b). This suggests that the relative weighting of reward information did not change over time and thus, can be assumed constant during each reversal.



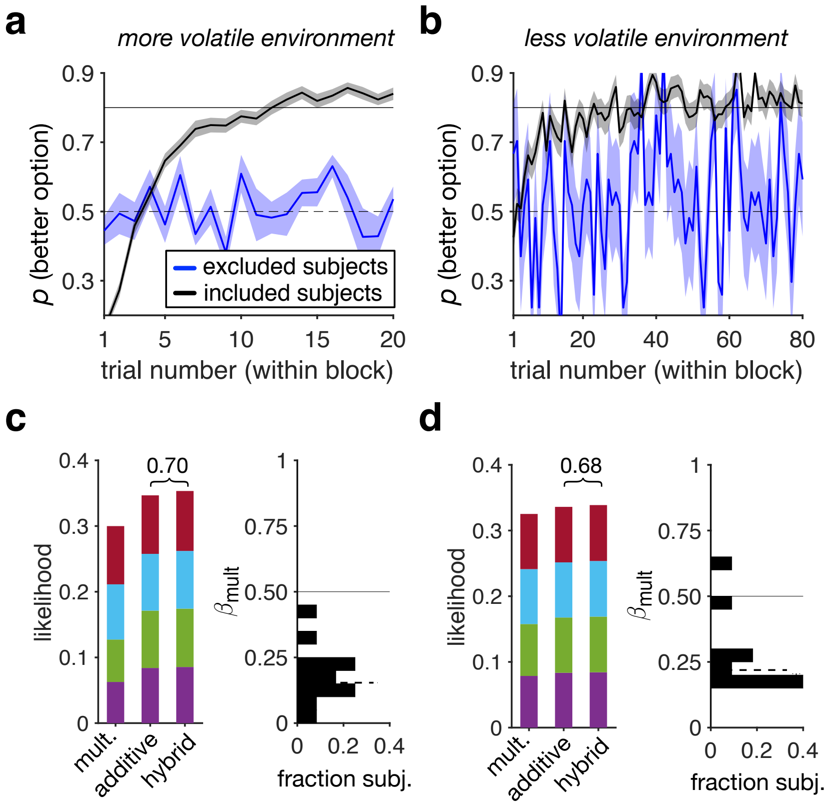
Supplementary Figure 1. Fitting procedure can correctly estimate model parameters used to generate data based on the hybrid model. (**a-d**) Plotted are the estimated relative weight of the multiplicative component (β_{mult}) as a function of actual β_{mult} and magnitude-to-probability weighting (β_m/β_p) used to generate the data, separately for gambling task (a) and three environments of the PRL task (b-d; stable: L = 200; less volatile: L = 80; more volatile: L = 20). The solid black curve indicates parameter values for which β_{EV} is equal to 0.5. Horizontal contours indicate that β_{mult} can be retrieved accurately and independently of magnitude-to-probability weighting. (**e-h**) Plotted are the estimated $\log(\beta_m/\beta_p)$ as a function of the relative weight of the multiplicative component (β_{mult}) and magnitude-to-probability weighting (β_m/β_p) used to generate the data. Except for β_{mult} close to 1, corresponding to a predominantly multiplicative model, magnitude-to-probability weighting can be retrieved accurately.



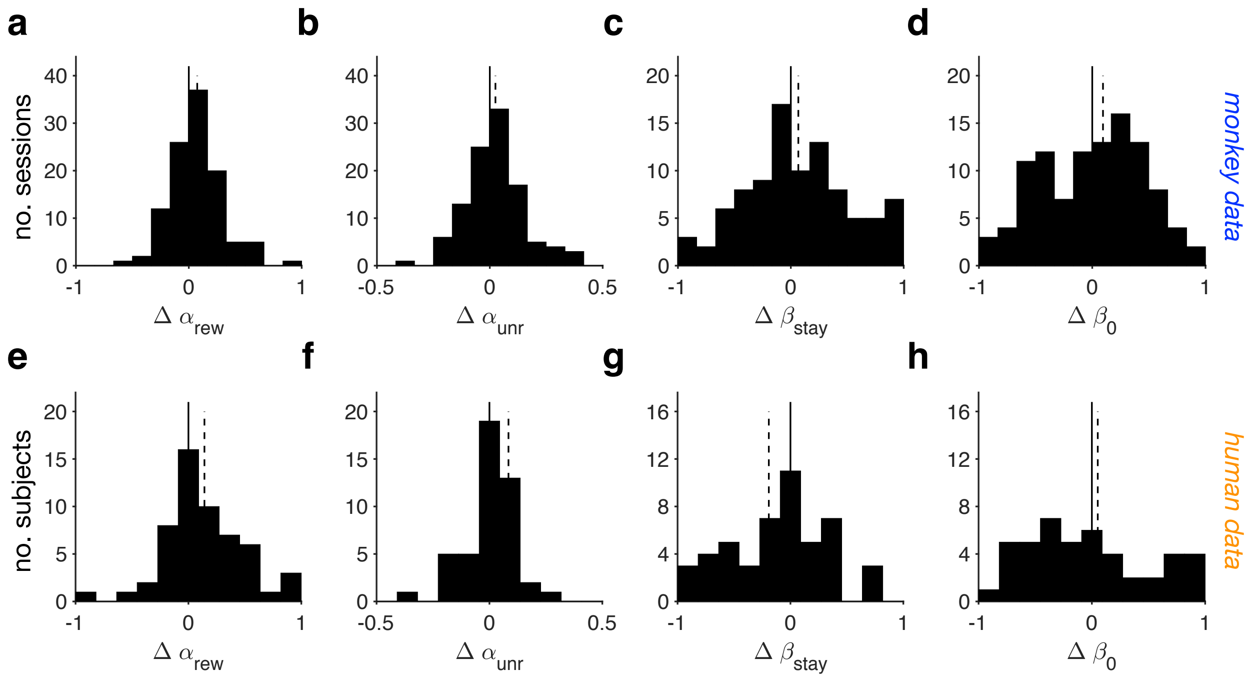
Supplementary Figure 2. Our method can identify the strategy most compatible with data generated using a hybrid model. (**a-d**) Likelihood of the hybrid model to be the identified model as a function of the relative weight of the multiplicative component (β_{mult}) and magnitude-to-probability weighting (β_m/β_p) used to generate the data, separately for the gambling task (a) and three environments of the PRL task (b-d; stable: $L = 200$; less volatile: $L = 80$; more volatile: $L = 20$). (**e-l**) The same as in a-d but showing the likelihood of the additive (e-h) or the multiplicative model (i-l) to be the identified model. (**m-p**) Plots show the differences between the likelihood of the additive and multiplicative models as a function of the parameters of the hybrid model used to generate the data (the same convention as in a-d). Positive (negative) values correspond to higher likelihood for the additive (multiplicative) model to be assigned as the correct model. The solid black curve indicates parameter values for which Δ likelihood is equal to 0, and the dashed horizontal line indicates $\beta_{mult} = 0.5$ above which Δ likelihood should be negative. Overall, our fitting method identifies the hybrid model as the most likely model followed by the additive and multiplicative when β_{mult} is close to 0 and 1, respectively. Moreover, we found some bias in identifying the more dominant component (additive vs. multiplicative) only in the PRL task for β_{mult} around 0.5, but this bias depended on magnitude-to-probability weighting. For very small β_m/β_p values, the model identification was biased toward the multiplicative strategy whereas there was a bias toward the additive strategy as β_m/β_p increased.



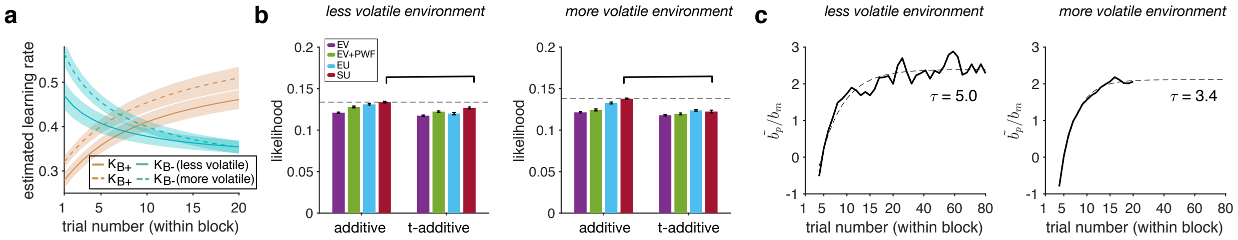
Supplementary Figure 3. Identification of different strategies for combination of reward information under risk and uncertainty using AIC. (**a**) Plotted are the average goodness-of-fit using AIC (across sessions) for monkeys during the gambling task (choice under risk, $N = 146$). Different colors indicate different models: expected value (EV), EV with probability weighting (EV+PW), expected utility (EU), and subjective utility (SU), used for the estimation of subjective value. (**b-c**) Same as in panel a but for monkeys in the stable (b) and volatile (c) environments of the mL task ($N = 316$). (**d-e**) Same as in panel b-c but for monkeys in the less volatile (d) and more volatile (e) environments of the PRL task ($N = 316$). (**f-j**) The same as in panel a-e but plotted are the average goodness-of-fit (AIC) across human participants during the three tasks (gambling: $N = 64$, mixed learning: $N = 46$, PRL: $N = 38$). Under risk, multiplicative models can explain choice behavior better for both monkeys and human participants, whereas additive models provide better fits to choice under uncertainty.



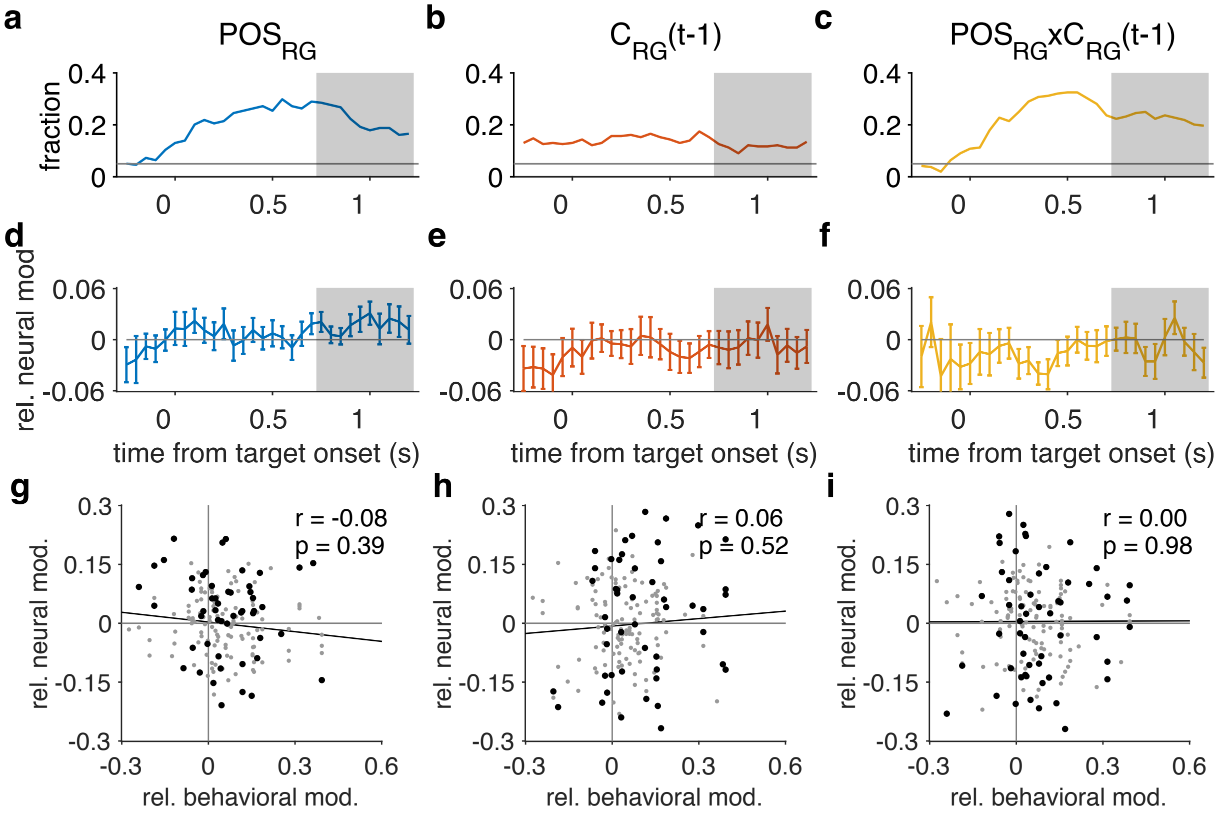
Supplementary Figure 4. Analysis of choice behavior of the included and excluded human participants in the PRL task (a-b) Plotted is probability of selecting the option with the higher probability of reward during each block of the PRL task, separately for the included (*N* = 38) and excluded human participants (*N* = 12) in the more volatile (a) and less volatile environments (b). The dashed and solid lines show chance level and the average probability of selecting the option with the higher probability of reward, respectively. Overall, excluded participants failed to learn reward probabilities associated with the two options. (**c-d**) Left panel: Likelihood of model adoption based on the BMS for the excluded participants. Right panel: Distribution of estimated values of β_{mult} using the hybrid models for the excluded participants. Conventions are the same as in Figure 2 and 3 of the main text. The medians of the distributions (dashed line) were significantly different from 0.5 (solid line) (two-sided Wilcoxon rank-sum test; more volatile environment: median±IQR: 0.15±0.09, $p = 7.8\times10^{-3}$, $d = 4.1$, $N = 12$, 95% CI = [0.27 0.42]; less volatile environment: median±IQR: 0.23±0.11, $p = 0.035$, $d = 1.8$, $N = 12$, 95% CI = [0.15 0.4]). There was no evidence that the excluded participants adopted strategies qualitatively differently than the participants included in our study.



Supplementary Figure 5. Behavioral adjustments in response to changes in volatility of the environment in the PRL task. (**a-d**) Plotted is the histogram of the difference in the estimated parameters of the simple additive model between the more and less volatile environments of the PRL task in monkeys ($N = 118$): learning rate on rewarded trials (a), learning rate on unrewarded trials (b), the tendency to select the target selected on the preceding trial (c), and the overall bias toward the green or red target (d). The solid and dashed lines show 0 and median, respectively. These results are based on session-by-session fit of the data. There was no significant change in any of the model parameters. (**e-h**) The same as in panels a-d but for human data ($N = 38$).



Supplementary Figure 6. Behavioral effects of volatility on learning and choice in the PRL task in monkeys. (**a**) Plotted are the average estimated effective learning rates over time on trials in which reward was assigned to the better (K_{B+}) and worse options (K_{B-}), separately for the less and more volatile environments. These estimates are obtained using the session-by-session fit of choice data assuming effective learning rates change over time according to an exponential function (N = 118). The shaded areas represent ±s.e.m. (**b**) Likelihood of different additive strategies (simple additive and time-dependent additive models) adopted by monkeys using the Bayesian model selection method. The bracket points to the best model in each strategy and the dashed line indicates the likelihood for the best overall model in each environment. SU models are the best simple additive and time-dependent additive (t-additive) models, and simple additive SU model is the better model in both environments. Overall, the simple additive models captured choice behavior better than the time-dependent ones. (**c**) Plotted are the ratio of the overall impact of reward probability to the weight of reward magnitude ($\tilde{b}_p/b_m$) in the less and more volatile environments. The dashed lines show the fit using an exponential function with the corresponding time constant reported in each panel.



Supplementary Figure 7. Other variables that are encoded in the response of dlPFC neurons but do not contribute to behavioral adjustments. (**a-c**) Plotted is the percentage of neurons that significantly encode the position of colors (a), the previous chosen color (b), and interaction of the position of colors and previous chosen color (c) ($N = 118$). (**d-f**) Plotted is the median of the relative neural modulation due to volatility (using estimated regression coefficients) across time for neurons that significantly encode the position of colors (d), the previous chosen color (e), and interaction of the position of colors and previous chosen color (f). Error bars show s.e.m. Gray background shows the period between 0.75s and 1.25s after target onset. These visualizations of the results of the regression were obtained with a sliding window of length 500ms. (**g-i**) Plotted is the change in dlPFC encoding of a given variable indicated in the top panels (relative neural modulation due to volatility) vs. relative behavioral modulation. Black dots indicate neurons that significantly encode the position of colors (g), the previous chosen color (h), and interaction of the position of colors and previous chosen color (i), and the grey dots indicate the rest of the neurons.