Neuron, Volume *73*

Supplemental Information

Rostrolateral Prefrontal Cortex and Individual Differences in Uncertainty-Driven Exploration

David Badre, Bradley B. Doll, Nicole M. Long, and Michael J. Frank

SUPPLEMENTAL INFORMATION

Supplemental Experimental Procedures

(a) *Computational Model Details and Procedures*

In addition to the components described in the introduction and Equations 1-4 in the main text, the full RT model included additional contributions to responding that were not a focus of the present experiment, for consistency with prior reports (but see alternative models). Thus, the full model estimates reaction time ($\hat{RT}$) on trial *t* as follows:

$$\hat{RT}(t) = K + \lambda RT(t-1) - Go(t) + NoGo(t) + \rho[\mu_{slow}(t) - \mu_{fast}(t)] + \nu[RT_{best} - RT_{avg}] + Explore(t) \quad (5)$$

where *K* is a free parameter capturing baseline response speed (irrespective of reward), λ reflects autocorrelation between the current and previous RT, and ν captures tendency to adapt RTs toward the single largest reward experienced thus far ("going for gold"). For details on these parameters, see Frank et al., 2009.

Go and NoGo learning reflect a striatal bias to speed responding as a function of positive RPE's and to slow responding as a function of negative RPEs. Evidence for speeding and slowing in the task is separately tracked:

$$\text{Go}(t) = \text{Go}(t\text{-}1) + \alpha_G \delta_+(t\text{-}1) \quad (6)$$

$$\text{NoGo}(t) = \text{NoGo}(t\text{-}1) + \alpha_N \delta_-(t\text{-}1) \quad (7)$$

where α_G and α_N are learning rates scaling the effects of positive (δ_+) and negative (δ_-) errors in expected value prediction *V* (i.e., positive and negative RPE). Go learning speeds RT, while NoGo learning slows it. This bias to speed and slow RTs as a function of positive and negative RPEs is adaptive in this task given that subjects tend to initially make relatively fast responses, and prior studies have found that these biases and model parameters are influenced by striatal dopaminergic manipulations and genetics (Moustafa et al, 2008; Frank et al, 2009). However, note that this approach does not consider when it is best to respond in a strategic manner, and in fact, it is not adaptive in environments where slow responses yield higher rewards (in which case, positive and negative RPE's will lead to maladaptive RT adjustments).

For the more strategic exploitative component, reward statistics were computed via Bayesian updating of "fast" or "slow" actions, as described in the main text. Fast or slow actions were classified based on whether they were faster or slower than the local average, which was computed as:

$$RT_{avg}(t) = RT_{avg}(t\text{-}1) + \alpha[RT(t\text{-}1) - RT_{avg}(t\text{-}1)] \quad (8)$$

However, fast and slow responses can be defined in other ways – such as based on whether the clock hand is in the first or second half of the clock face – and outcomes from the model are the same. (The use of an adaptive version of the boundary is more general and would allow the algorithm to converge to an appropriate RT even if reward functions are non-monotonic).

Free parameters were estimated for each subject via the Simplex method as those minimizing the sum of squared error between predicted and observed RTs. Multiple starting points were used for each optimization process to reduce the likelihood of local minima. All parameters were free to vary for each participant, with the exception of α used in the expected value (*V*) update, which was set to 0.1 for all participants to prevent model degeneracy (Frank et al., 2009).

For other details regarding the primary continuous RT model, including alternative models that provide poorer behavioral fits to the data please see Frank et al (2009). Note that among the alternative models tested in that paper is a Kalman filter model in which the mean expected reward values and their uncertainties are estimated with Normal distributions, rather than the beta distributions used here. However, the variance (uncertainty) in the Kalman filter tends to be overly dominated by the first trial: subjects are given no information about the number of possible points that they might gain, leading to large variance in initial estimates, which then declines more dramatically after a few trials when rewards are experienced than does the variance in the probability distributions for the beta priors. This means that the neural estimate of relative uncertainty would largely reflect contributions of a very few number of trials using this model, making it inappropriate for estimating fMRI data in the present data set. For the "RT swing model", the identical procedure was used, except parameters were optimized to predict the change in RT from one trial to the next, rather than the raw RT. For the "sticky choice" simulations, we estimated the effect of not just the prior trials' RT (with parameter λ), but instead a decaying function of previous RT's. Specifically, we replaced λ RT(t-1) with λ sticky(t), where sticky(t) = RT(t-1) + *d sticky*(t-1),and $0<d<1$ is a decay parameter influencing the degree to which prior RT's continue to affect current RT's.

For the simplified two-alternative choice models, we used a softmax function to predict the probability of a fast or slow response:

$$P_{fast}(t) = \frac{1}{1 + e^{\beta[b+\rho(\mu_{fast}-\mu_{slow})+\epsilon(\sigma_{fast}-\sigma_{slow})]}} \quad (9)$$

where β is the softmax gain parameter, *b* is a bias parameter estimating the degree to which an individual is more or less likely to respond fast or slow independent of reward history (analogous to *K* in the RT model), and the other parameters are identical to those in the RT model. In the Q learning version of this model, the means of the beta distributions were replaced with Q values for fast and slow actions, each of which were updated with an additional learning rate parameter α:

$$Q_{s,a}(t) = Q_{s,a}(t-1) + \alpha[Rew(t) - Q_{s,a}(t-1)] \quad (10)$$

where *s* represents the state (a given clock-face) and *a* the action (fast or slow). Thus the Q learning version allowed the expected values to be updated as a function of prediction error without requiring updating to proceed in a Bayesian manner, but we still allowed uncertainty as derived from the beta distributions to guide exploration. In both models using beta distributions or Q values we optimized parameters by maximizing the log likelihood of each participant's trial-by-trial sequence of responses (binarized as fast or slow), and compared model fits (using likelihood ratio tests) between models that incorporated ε versus those that fixed ε=0. In follow up simulations we also included a sticky choice parameter λ analogous to that in the standard model (i.e., increasing the probability of selecting the same action as the prior trial) by incorporating this into the softmax function (e.g., Schonberg et al, 2007).

(b) *fMRI procedures and analysis*

Whole-brain imaging was performed on a Siemens 3T TIM Trio MRI system. Functional images were acquired using a gradient-echo echo-planar sequence (TR = 2 s; TE = 30 ms; flip angle = 90°; 40 axial slices, 3 x 3 x 3 mm). After the five functional runs, high-resolution T1-weighted (MPRAGE) anatomical images were collected for visualization (TR = 1900 ms; TE = 2.98 s; flip angle = 9°; 160 sagittal slices, 1 x 1 x 1 mm). Head motion was restricted using firm padding that surrounded

the head. Visual stimuli were projected onto a screen and viewed through a mirror attached to a matrix thirty-two channel head coil. Responses were registered on a Mag Design and Engineering MRI-compatible four-button response pad.

Preprocessing and data analysis were performed using SPM2 (http://www.fil.ion.ucl.ac.uk/spm/). Following quality assurance procedures to assess outliers or artifacts in volume and slice-to-slice variance in the global signal, functional images were corrected for differences in slice acquisition timing by resampling all slices in time to match the first slice. Images were then motion corrected across all runs (using b-spline interpolation). Functional data were then normalized based on MNI stereotaxic space using a 12-parameter affine transformation along with a nonlinear transformation using cosine basis functions. Images were resampled into 2-mm cubic voxels and then spatially smoothed with an 8-mm FWHM isotropic Gaussian kernel.

Data analysis was conducted under the assumptions of the general linear model as implemented in SPM2. All regressors were generated by convolving event epochs with a canonical hemodynamic response function and its temporal derivative. Separate event-related regressors were generated for the onset of stimulus and reward events. Two models were constructed to examine relative uncertainty effects. In the first model, relative uncertainty was included as a parametric modulator in association with stimulus onset followed by mean uncertainty. The relative uncertainty regressor was computed as the absolute value of the difference in the standard deviations of the expected value distributions associated with fast and slow responses on each trial (relative uncertainty(t) = $|\sigma slow(t) - \sigma fast(t)|$). The mean uncertainty regressor reflects changes in the magnitude of uncertainty across all responses and was simply computed as the mean of the standard deviations of the expected value distributions associated with fast and slow responses on each trial (mean uncertainty(t) = $[\sigma slow(t)+\sigma fast(t)]/2$). In the second model, mean uncertainty was entered prior to relative uncertainty. The order of the parametric regressors affects the way that shared variance is explained between them, such that (as implemented in SPM) the first parametric includes both unique and shared variance and the second only unique. Thus, in the first model, explained variance by the mean uncertainty regressor is that which goes above and beyond that explained by relative uncertainty. And in the second model, variance explained by the relative uncertainty regressor is that which goes above and beyond mean uncertainty. In both models, additional parametric regressors reflecting positive RPE ($\delta+$) and negative RPE ($\delta-$), and overall RT were modeled at reward onset to account for variance due to these factors.

Four separate additional GLM models were constructed in order to test the hypothesis that RLPFC tracks the value of the unchosen option. These GLMs were constructed identically to the second relative uncertainty GLM described above, except that the relative uncertainty regressor was replaced in separate models by (a) μ of the unchosen option (μunchosen), (b) the difference in μ between the unchosen and chosen option (μunchosen $-$ μchosen), (c) the log ratio of μunchosen versus μchosen (log[μunchosen/μchosen]), and (d) "exploration against the odds" (μbest $-$ μchosen).

Statistical effects were estimated using a subject-specific fixed-effects model, with session specific effects and low-frequency signal components (< .01 Hz) treated as confounds. Linear contrasts of the whole brain were used to obtain subject-specific estimates for each effect. These estimates were entered into a second-level analysis treating subjects as a random effect, using a one-sample t-test against a contrast value of zero at each voxel. Voxel-based group effects from whole brain analysis were considered reliable to the extent that they survived a family-wise error (FWE) corrected threshold of $p < .05$ at the cluster level. For smaller structures, like the nucleus accumbens, use of a cluster level correction can be inappropriate. Thus, for this structure of a priori interest, we used a whole brain voxel level false discovery rate (FDR) correction of $p < .05$. Group contrasts were rendered on an MNI canonical brain that underwent cortical "inflation" using FreeSurfer (CorTechs Labs, Inc.) (Dale et al., 1999; Fischl et al., 1999).

Whole brain analyses were complemented by region of interest (ROI) analyses to test predicted effects in a priori hypothesized regions. Functionally defined ROIs were chosen based on all significant

voxels within an 8-mm radius of a chosen maximum from the unbiased contrast of all stimulus and feedback onsets versus fixation. For parametric regressors, the average beta for that regressor among voxels in the ROI was calculated as an estimate of average effect size. The resultant data were subjected to repeated-measures analyses of variance and t tests as noted in the results.

Supplemental Results

(c) *Supplemental Model Analysis*

For explorers ($\varepsilon > 0$), the mean sum of squared error (SSE) between predicted and actual response times across all trials was 1.14×10^8 (standard error = 1.6×10^7). This is in the same range as that previously reported for the best-fitting model across a sample of 70 participants with similar demographics in Frank et al (2009), adjusting for the fact that there were twice as many trials here for fMRI (i.e. the error per trial is comparable). As reported previously, this represented an improvement in fit compared to a model that assumes no uncertainty-driven exploration ($\varepsilon = 0$; mean $\Delta SSE = 5.54 \times 10^5$, standard error = 2.8×10^5). The previous study also reported that depending on one's genotype, the inclusion of an uncertainty exploration parameter improved model fit when also penalizing for the added model complexity using Aikake's Information Criterion (AIC). In this (far smaller) sample, this statistic was not reliable (but see below for supplemental analysis); however, the improvement in AIC was nevertheless correlated with fitted ε value across subjects ($r=0.68$, $p < .005$), and several analyses in the main text demonstrated that the improvement of fit was reliable in the explorer sub-group, and in some models across the entire group. Moreover, fMRI data reported in the main text revealed that those with positive ε values reliably exhibited neural activity in frontopolar cortex that tracked relative uncertainty.

(d) *fMRI analysis of relative uncertainty in the first half of blocks*

A task where the primary behavioral measure is RT can conceivably be affected by session and block-related confounds, like fatigue and boredom, that will be more likely at the end of blocks. Moreover, in the current task, the contribution of relative uncertainty could potentially be greater at the beginning of blocks when participants are generally more uncertain. Thus, to establish that the reported effects of relative uncertainty hold during the first half of a block, we re-ran the GLM in which mean uncertainty is entered as a parametric modulator of stimulus onset before relative uncertainty. The only difference from the version reported in the text was that we restricted analysis to the first half of experimental trials in the block.

This analysis yielded results fully consistent with those reported in the main text. In particular, we found an effect ($p < .05$ [FWE cluster level]) of relative uncertainty only in the explore participants in dorsal (XYZ = 32 53 18) and ventral RLPFC (XYZ = 40 58 -8), along with SPL (-16 -70 58) and cerebellum (XYZ = 42 -62 -34). There was no effect in the non-explore participants. Thus, the reported results were not due to a fatigue or boredom-related effect more evident at the end of blocks. Also, beyond this analysis, the effect of mean uncertainty controlled for uncertainty related monotonic declines over the course of a run, and so provide additional assurance that these low frequency components are not driving the relative uncertainty effect.

(e) *Relative uncertainty effects in prior definitions of RLPFC*

As noted in the Introduction, RLPFC has been previously associated with exploration (e.g., Daw et al., 2006) and branching comparisons, such as in tracking the value of unchosen options (Boorman et al., 2009). However, the definition of RLPFC is not always the same across studies, and so it is important to establish that putatively similar effects are indeed in the same region of cortex. Thus, to draw a tighter link with this prior work, we sought to directly test the effect of relative uncertainty in ROIs defined from these studies.

First, we tested the effect of relative uncertainty in right (XYZ = 27 57 6) and left (XYZ = -27 48 4) RLPFC ROIs defined based on the coordinates reported in Daw et al., (2006). Consistent with the

present results, there was a reliable effect relative uncertainty in the explore participants in right RLPFC ($t(7) = 2.5$, $p < .05$). There was no effect of relative uncertainty in right RLPFC non-explore participants or in left RLPFC in either group of participants (t's < .7). We also tested ROIs defined in left (XYZ = -29 -33 45) and right (XYZ = 39 -36 42) intraparietal sulcus (IPS) using peak coordinates reported in Daw et al. (2006). Consistent with the whole brain analysis reported in the main paper, there was no effect of relative uncertainty in either IPS ROI in explore or non-explore participants (ts < 1.8). Thus, we found a reliable parametric effect of relative uncertainty in the same right RLPFC region highlighted by Daw et al. (2006) in association with exploration.

Next, we tested the ROIs in left (XYZ = -34 56 -8) and right (XYZ = 36 54 0) RLPFC and mid-IPS (left: -32 -60 52; right: 50 -46 46) identified in association with tracking the value of the unchosen option during decision making (Boorman et al., 2009). Again, this result located reliable effects of relative uncertainty in the right RLPFC ROI in whole group ($t(14) = 2.3$, $p < .05$) and in the explore participants ($t(7) = 3.5$, $p = .01$). But, no effect of relative uncertainty in the non-explore participants in right RLPFC ($t = .2$) or in the other ROIs tested.

(f) *Branching and the expected reward of the unchosen option*

One potential alternative hypothesis for the function of RLPFC during exploration is that it reflects maintenance of the mean reward probability of the unchosen option on every trial, rather than the relative uncertainty about it. Maintenance of pending states or courses of action, also termed "branching", has been previously associated with RLPFC (Koechlin et al., 1999). Moreover, as mentioned earlier, a prior study demonstrated that RLPFC tracks the unchosen reward probability in the service of future choices (Boorman et al., 2009). In order to test this hypothesis in the current task, we conducted a series of analyses using the trial-to-trial estimates of the mean expected value of the unchosen option in our fMRI analysis. Specifically, in separate models, we replaced relative uncertainty with regressors based on (a) the mean probability of a positive RPE for the unchosen option (μ_{unchosen}), (b) the relative difference between the means of the unchosen and chosen option ($\mu_{\text{unchosen}} - \mu_{\text{chosen}}$), (c) the log ratio of the means of the two options ($\log[\mu_{\text{unchosen}}/\mu_{\text{chosen}}]$; (Boorman et al., 2009), and (d) "exploration against the odds" ($\mu_{\text{best}} - \mu_{\text{chosen}}$; Daw et al., 2006). However, these analyses failed to locate activation in RLPFC in association with these parametric functions, including at reduced statistical thresholds. These null results should not be interpreted as evidence against a general branching mechanism for RLPFC in all task contexts. However, it does suggest that, in the present task, consideration of alternative choices is better accounted for by relative uncertainty about the values – the information to be gained by exploring – rather than as a function of expected rewards among unchosen options.

(g) Analysis of Reward Prediction Error

RPE signals are hypothesized to underlie both exploration and exploitation decisions. Consistent with prior studies of reward and reinforcement learning (Gershman et al., 2009; McClure et al., 2003; O'Doherty et al., 2003; Rutledge et al., 2010; Badre and Frank, In Press), estimates of RPE from our model were associated with ventral striatal, and medial and lateral PFC regions (Supp. Fig. 1). We first estimated signal change related to the onset of feedback (i.e., the presentation of how many points were won on each trial) versus baseline (Supp. Fig. 1a). This contrast yielded activation ($p < .001$ [FWE cluster level]) in right insula (XYZ = 34 20 2; 30 10 -4), left and right ventral and lateral occipital cortex (XYZ = 14 -80 -18; -38 -60 -28), bilateral posterior parietal cortex (XYZ = 44 -56 50; -44 -50 50), dorsomedial frontal cortex (XYZ = 12 16 62), and right lateral frontal cortex (XYZ = 26 52 -18; 34 10 64).

We next assessed positive RPE (when rewards are better than expected) and negative RPE (when rewards are worse) separately in the fMRI data in order to distinguish their respective contributions to signal variance beyond that associated with feedback onset (Supp. Fig. 1b). Consistent with prior fMRI studies of reinforcement learning, positive RPE produced activation in bilateral ventral striatum (XYZ = 16 0 -14; XYZ = -14 4 -4; $p < .001$ [FWE cluster level]). This finding is consistent with models and data showing that striatal dopaminergic manipulations affect the degree to which individuals learn from positive vs negative RPE's in this task (Moustafa et al, 2008; Frank et al, 2009). Positive RPE activation ($p < .001$ [FWE cluster level]) was also observed in a network of neocortical regions, including bilateral rostral inferior frontal gyrus (XYZ = 44 44 -14; XYZ = -44 42 -12), ventral occipital (XYZ = 18 -94 -26), and posterior parietal cortex (XYZ = 42 -60 42; XYZ = -48 -60 54).

Analysis of negative RPE versus baseline did not yield reliable effects at corrected thresholds. Direct contrast of positive minus negative RPE yielded a single focus in ventral striatum, specifically in the left nucleus accumbens (XYZ = -12 2 -2, $p < .05$ [FDR voxel level]). Beyond the striatum, positive versus negative RPE activated a similar neocortical network as that observed for positive RPE versus baseline ($p < .001$, [FWE cluster level]).

Supplemental Discussion

(h) *Comparison of Relative Uncertainty with other forms of uncertainty*

Relative uncertainty in the present study refers to a specific form of uncertainty, namely uncertainty about the probability of an action yielding a positive RPE. However, there are other forms of uncertainty during decision making in this task which might affect behavior; though we would argue that these are not critical for the relative uncertainty computation that underlies exploration.
Beyond the calculation of mean expected value, it is also possible that participants routinely track ambiguous (uncertain) choices to weigh directly against known risks. Tasks that require participants to make such direct choices between ambiguous versus risky options associate DLPFC with the choice of ambiguous options over risky ones (Huettel et al., 2006; Payzan-LeNestour and Bossaerts, 2011). Thus, a neural representation of the individual uncertainty associated with each option (akin to mean uncertainty in DLPFC) may also influence response choices in the task directly, though the contribution of such effects were not estimated in the model. Future work will be needed to specify and expand on mean uncertainty representations in prefrontal cortex.

Beyond uncertainty aversion, aversion to known risks (akin to outcome uncertainty, see below) might play a role in participant choices. We have examined this sort of risk aversion in behavioral analysis by calculating the extent to which participants adjust their RTs in the CEV condition relative to CEVR (Moustafa et al, 2008; Frank et al, 2009; Strauss et al, 2011). Both cases have equal and constant expected values for all RTs, but in one condition reward probability goes down with time while magnitude goes up, and in the other it is the opposite. A participant's systematic tendency to respond slower in CEVR would be indicative of risk aversion, because it suggests a preference for a high probability of a small gain over a lower probability of a large gain. It was previously reported that this form of risk aversion is subject to additive influences of both striatal and prefrontal dopaminergic genetic variants (Frank et al, 2009) and dopaminergic manipulation (Moustafa et al, 2008). However, in present sample, we did not locate reliable group differences between CEVR and CEV at later trials.
Payzan and Bossaerts (2011) discussed three forms of uncertainty and their contributions to reinforcement learning. The first is irreducible uncertainty, and corresponds to known risk about the outcome in any given trial. This form of uncertainty is not relevant for exploration: once a learner is confident that a particular choice is associated with a certain level of value there is no information to be gained by exploring that option. The second form is termed estimation uncertainty or ambiguity and quantifies ignorance about the reinforcement statistics in terms of the variance. The third form of uncertainty is unexpected uncertainty (Yu & Dayan, 2005) which reflects the learner's belief of the probability that the action-outcome contingencies have suddenly changed. Under this scenario, unexpected outcomes can be taken as an indicator that the reward distributions should be re-initialized, and therefore exploration should begin anew. We have eliminated this aspect from our task paradigm by keeping the task contingencies stationary within a block of trials, and then reinitializing them when a new clock-face is presented in a new block. Thus we fit the data with a model that reinitializes the distributions to be uniform at the outset of each block, encouraging renewed exploration, without having to model participants' belief that the contingencies may have changed.

Finally, another form of uncertainty, not tested here, might arise from decision-level conflict. When the learned value of different actions is very similar, this may produce conflict (which often is associated with slowed RTs), but it would not produce a change in RT in the direction of greater estimation uncertainty (which would sometimes be faster and sometimes slower, depending on which action had the most uncertain reward statistics). The model also uses the relative difference in the mean reward estimates to drive exploitation, and when these means are most similar, the model would just predict an

intermediate RT. If there was a further bias to slow down due to this conflict in mean estimates ('conflict-induced slowing'), this would be orthogonal to the RT swings that we model by relative uncertainty. However, it is unclear how this type of conflict based uncertainty would predict the directional RT swings that would correlate with those associated with relative uncertainty.

(i) *Neural Correlates of Mean Uncertainty*

Growing evidence has suggested that the computation of higher order relations in RLPFC arises from its position at the apex of a functional gradient arrayed along the rostro-caudal axis of lateral frontal cortex (Badre, 2008; Koechlin and Summerfield, 2007). From this perspective, rostral regions maintain more abstract, higher order representations than caudal frontal cortex, with the most abstract representations being coded in RLPFC. Between-level interactions may exist such that higher order relations in rostral regions are computed over more concrete representations maintained in caudal regions (Christoff and Gabrieli, 2000; Koechlin et al., 2003). Potentially consistent with this perspective, we observed that the absolute level of uncertainty, as indexed by mean uncertainty, activated a broader area of PFC including more caudal DLPFC, whereas the effect of relative uncertainty was greater in RLPFC than DLPFC. One interpretation of this result is that neurons in DLPFC may code for the probability of the mean expected value, as observed in non-human primates (Kim et al., 2008; Kim et al., 2009), either directly or in conjunction with other regions.

It should be noted, however, that because mean uncertainty is more monotonic, generally decreasing across a block of trials, we cannot rule out the possibility that the DLPFC mean-uncertainty effects are due to other aspects of cognitive control that decline with time on task or practice (e.g., Raichle et al., 1994). And, indeed, the effect of mean uncertainty was not focal, even after controlling for relative uncertainty. Rather, correlates were located in a large neocortical network inclusive of occipital and parietal cortex and RLPFC. However, mean uncertainty declined only within a block of trials and then rose again at the outset of the next block, so this alternative interpretation should not be confused with a global practice effect in the task, fatigue, or other monotonically declining functions occurring over the entire experimental session. Moreover, as variance due to mean uncertainty was removed prior to that associated with relative uncertainty, the effect of relative uncertainty in RLPFC is not due changes in mean uncertainty, irrespective of the source of this latter effect.

Unlike the RLPFC relative uncertainty effects, the effect of mean uncertainty in DLPFC and other regions was evident in both explore and non-explore participants. Thus, tracking the uncertainty associated with mean expected value of an option may be advantageous, even if it is not being used to compute relative uncertainty. In the present model, because it is Bayesian, the uncertainty associated with each option is directly related to the effective learning rate. More specifically, as learning progresses over the session, RPE's are weighted to a different degree in updating the probabilistic value estimates of an option (i.e., evidence under conditions of high uncertainty affects the mean more than under conditions of low uncertainty). The neural mechanisms by which the brain integrates new evidence into the mean expected value during learning remain to be specified. But, it is conceivable that uncertainty estimates maintained by right DLPFC contribute to this integration process.

(j**) Supplemental References**

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. Trends Cogn Sci *12*, 193-200.

Badre, D., and Frank, M.J. (In Press). Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 2: Evidence from fMRI. Cerebral Cortex.

Boorman, E.D., Behrens, T.E., Woolrich, M.W., and Rushworth, M.F. (2009). How green is the grass

on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. Neuron *62*, 733-743.
Christoff, K., and Gabrieli, J.D.E. (2000). The frontopolar cortex and human cognition: Evidence for a rostrocaudal hierarchal organization within the human prefrontal cortex. Psychobiology *28*, 168-186.
Dale, A.M., Fischl, B., and Sereno, M.I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. NeuroImage *9*, 179-194.
Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., and Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. Nature *441*, 876-879.
Fischl, B., Sereno, M.I., and Dale, A.M. (1999). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. Neuroimage *9*, 195-207.
Frank, M.J., Doll, B.B., Oas-Terpstra, J., and Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci *12*, 1062-1068.
Gershman, S.J., Pesaran, B., and Daw, N.D. (2009). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. J Neurosci 29, 13524-13531.
Huettel, S.A., Stowe, C.J., Gordon, E.M., Warner, B.T., and Platt, M.L. (2006). Neural signatures of economic preferences for risk and ambiguity. Neuron *49*, 765-775.
Koechlin, E., Ody, C., and Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. Science *302*, 1181-1185.
Koechlin, E., and Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. Trends Cogn Sci *11*, 229-235.
Kim, S., Hwang, J., and Lee, D. (2008). Prefrontal coding of temporally discounted values during intertemporal choice. Neuron *59*, 161-172.
Kim, S., Hwang, J., Seo, H., and Lee, D. (2009). Valuation of uncertain and delayed rewards in primate prefrontal cortex. Neural Netw *22*, 294-304.
McClure, S.M., Berns, G.S., and Montague, P.R. (2003). Temporal prediction errors in a passive learning task activate human striatum. Neuron 38, 339-346.
Moustafa, A.A., Cohen, M.X., Sherman, S.J., and Frank, M.J. (2008). A role for dopamine in temporal decision making and reward maximization in parkinsonism. J Neurosci *28*, 12294-12304.
O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., and Dolan, R.J. (2003). Temporal difference models and reward-related learning in the human brain. Neuron 38, 329-337.
Payzan-LeNestour, E., and Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. PLoS Comput Biol *7*, e1001048.
Raichle, M.E., Fiez, J.A., Videen, T.O., MacLeod, A.M., Pardo, J.V., Fox, P.T., and Petersen, S.E. (1994). Practice-related changes in human brain functional anatomy during nonmotor learning. Cereb Cortex *4*, 8-26.
Rutledge, R.B., Dean, M., Caplin, A., and Glimcher, P.W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. J Neurosci 30, 13525-13536.
Schonberg, T., Daw, N.D., Joel, D., and O'Doherty, J.P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J Neurosci *27*, 12860-12867.
Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. Neuron. *46*, 681-692.

Supplemental Tables

(k) **Table S1.** Mean ε (SEM) values across alternate models in explorers and non-explorers as defined by primary model.

Model	Explorers (n = 8)	Non-Explorers (n = 7)
Primary	1657 (468)	0
Explore trials only	12746 (12520)	-3985(1113)
Sticky choice	889 (924)	-2296 (591)
RT swing	15693 (3907)	9065 (2362)
Softmax (no sticky choice)	4.57 (2.1)	-5.0 (1.8)
Softmax + sticky choice	8.4 (2.2)	-0.34 (1.98)
Softmax + sticky choice + Q-learning	5.8 (1.7)	1.3 (0.7)
Softmax + Q-learning (no sticky choice)	3.5 (1.26)	-0.41 (0.37)

(Note that ε values are on different scales across model variants, especially as they pertain to different outcome measures: RTs, RT swings, softmax choices.)

(l) **Table S2.** Activation foci from analyses of relative uncertainty

Region	Stereotaxic Coordinates X	Y	Z	~Brodmann's Area	Peak Z
Primary Model – All Participants					
Right RLPFC	36	56	-8	10	4.3
	32	66	-8	10	3.9
	20	68	-14	10	3.8
Right SPL	8	-70	62	7	3.7
	16	-66	46	7	3.3
Bilateral Occipital	-12	-96	0	17,18	6.1
	16	-96	8	17,18	5.8
	-6	-92	-12	17,18	5.4
Right DLPFC	30	30	28	46,9	4.4
	40	34	30	46,9	3.7
	40	26	36	46,9	3.6
Right IPS	42	-54	38	39,40	4.3
	48	-52	32	39,40	4.3
	42	-56	58	39,40	3.2
Right Operculum	30	24	-15	47	3.2
Right Cerebellum	44	-56	-34		3.8
Left Cerebellum	-44	-76	-24		4.4
Primary Model – Explore Participants Only					
Right RLPFC (dorsal)		24	48	20	10,46	4.3
	30	52	16	10,46	4.0
	18	40	22	10,46	3.9
Right RLPFC (ventral)	40	60	-10	10	4.1
	30	52	-14	10	4.1
	42	46	-2	10	3.8
Bilateral Occipital	-8	-96	2	17,18	5.0
	8	-90	-4	17,18	4.9
	-6	-90	-8	17,18	4.9
Right Cerebellum	28	-66	-28		4.2
Left Cerebellum	-28	-64	-32		3.7
Primary Model – Explore Participants Only (Controlled for Mean Uncertainty)					
Right RLPFC (dorsal)		44	42	28	10,46	4.7
	22	56	26	10,46	4.1
	26	52	16	10,46	4.0
Right RLPFC (ventral)	30	52	-14	10	4.2
	36	56	-10	10	4.2
Left SPL	-8	-62	66	7	4.3
	-16	-70	62	7	4.1
	-24	-68	68	7	3.5
Explore Trials Only Model - Explore Participants Only						Right RLPFC
30	-56	12		10	4.3	
Right IPS	36	-46	56	40	3.9	
	40	-54	58	40	3.7	
	44	-32	56	40	3.6	
Sticky Choice Model - Explore Participants Only					
Right RLPFC	26	52	16	10,46	4.2
	22	56	24	10,46	3.6
Left SPL	-6	-60	60	7	4.5
	-12	-70	72	7	4.3
	-16	-70	60	7	3.7
Right SPL	12	-64	64	7	4.1
Left Occipital	-36	-88	-16	18,19	4.8
Right Cerebellum	50	-56	-40		4.1
Softmax Model - Explore Participants Only					
Right RLPFC (dorsal)		44	42	28	10,46	4.7
	24	50	18	10,46	3.8
	34	52	16	10,46	3.6
Right RLPFC (ventral)	36	56	-10	10	3.9
Left SPL	-8	-64	66	7	5.0

Supplemental Figure

(m) **Figure S1.** *Effects of feedback and positive reward prediction error.* (a) Feedback onset activated in right ventral striatum, insula, and lateral and medial prefrontal cortex. (b) Positive reward prediction error accounted for variance beyond that associated with feedback onset in bilateral ventral striatum, insula, and lateral and medial PFC.