Dissociable Neural Correlates of Uncertainty Underlie Different Exploration Strategies

Momchil S. Tomov[1,2] , Van Q. Truong[2] , Rohan A. Hundia[2], Samuel J. Gershman[2]

[1]*Program in Neuroscience, Harvard Medical School, Boston, MA 02115, USA*
[2]*Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138, USA*

Running title: Uncertainty and Exploration in the Brain

Corresponding Author:
Momchil S. Tomov
52 Oxford St., Room 295.06
Cambridge, MA 02138
e-mail: mtomov@g.harvard.edu

Number of pages: 34

Number of figures: 7

Number of tables: 6

Number of words Abstract: 206

Number of words Introduction: 607

Number of words Discussion: 719

Conflict of Interest: The authors declare no competing financial interests.

Acknowledgements: This research was supported by the Office of Naval Research Science of Autonomy program (N00014-17-1-2984) and by the National Institutes of Health, award number 1R01MH109177. This work involved the use of instrumentation supported by the NIH Shared Instrumentation Grant Program grant number S10OD020039. This work also received funding from the Irene and Eric Simon Brain Research Foundation. We acknowledge the University of Minnesota Center for Magnetic Resonance Research for use of the multiband-EPI pulse sequences. We are grateful to Wouter Kool and Eric Schulz for giving us feedback on the manuscript.

Summary

Most real-world decisions involve a delicate balance between exploring unfamiliar alternatives and committing to the best known option. Uncertainty lies at the core of this "explore-exploit" dilemma, for if all options were perfectly known, there would be no need to explore. Yet despite the prominent role of uncertainty-guided exploration in decision making, evidence for its neural implementation is still sparse. We investigated this question with model-based fMRI (n = 31) using a two-armed bandit task that independently manipulates two forms of uncertainty underlying different exploration strategies. The relative uncertainty between the two options was encoded in right rostrolateral prefrontal cortex and drove directed exploration, a strategy that adds an uncertainty bonus to each option. The total uncertainty across the two options was encoded in left anterior insula and drove random exploration, a strategy that increases choice stochasticity in proportion to total uncertainty. The subjective estimates of uncertainty from both regions were predictive of subject choices. In addition, the encoding of total uncertainty in anterior insula across subjects predicted differences in the extent to which they employed random exploration. These results are consistent with a hybrid computational architecture in which different uncertainty computations are performed separately and then combined by downstream decision circuits to compute choice.

In Brief

Tomov et al. show that humans explore unfamiliar options using two strategies driven by distinct uncertainty computations. Relative uncertainty is encoded in right rostrolateral prefrontal cortex and guides directed exploration, while total uncertainty is encoded in left anterior insula and guides random exploration.

Highlights

- People use directed and random exploration during reward-based learning
- Relative uncertainty is encoded in right rostrolateral prefrontal cortex and drives directed exploration
- Total uncertainty is encoded in left anterior insula and drives random exploration
- Individual differences in the encoding of total uncertainty predict differences in exploration strategy

Introduction

For every decision that we make, we have to choose between the best option that we know so far (exploitation), or a less familiar option that could be even better (exploration). This "explore-exploit" dilemma appears at all levels of decision making, ranging from the mundane (Do I go to my favorite restaurant, or try a new one?) to the momentous (Do I marry my current partner, or try a new one?). Despite its ubiquity in everyday life, little is known about how the brain handles the exploration-exploitation trade-off. Intuitively, either extreme is undesirable: an agent that solely exploits will never adapt to changes in the environment (or even learn which options are good in the first place), while an agent that solely explores will never reap the fruits of that exploration. Yet striking the perfect balance is computationally intractable beyond the simplest examples, and hence humans and animals must adopt various heuristics (Cohen et al., 2007; Mehlhorn et al., 2015).

Earlier research suggested that people choose options in proportion to their expected values (Daw et al., 2006; Yechiam and Busemeyer, 2005), a strategy known as softmax exploration that is closely related to other psychological phenomena such as probability matching (for a detailed review, see Schulz and Gershman (in press)). Later studies showed that people explore in a more sophisticated manner, using uncertainty to guide their choices towards more promising options (Wilson et al., 2014; Gershman, 2018a). These strategies are more adaptive in nonstationary environments and come in two distinct flavors: *directed* and *random* exploration strategies.

Directed exploration strategies optimistically prefer options with less certain outcomes by adding an uncertainty bonus to their subjective values. For example, for your next meal, you might forego your favorite restaurant for a new one that just opened down the street, and you might even go there several times until you are certain it is no better than your favorite. Thus while softmax exploration is sensitive to the *relative value* of each option, preferring options with higher payoffs, directed exploration is additionally sensitive to the *relative uncertainty* of each option, preferring options with more uncertainty as they hold greater potential for gain. Directed exploration is closely related to the phenomenon of risk-seeking (Hertwig et al., 2004; Weber et al., 2004) and has strong empirical support (Frank et al., 2009; Speekenbrink and Konstantinidis, 2015; Dezza et al., 2017).

In reinforcement learning, directed exploration is often implemented via the upper confidence bound (UCB) algorithm (Auer, 2002), in which the bonus is the one-sided confidence interval of the expected value:

$$a_t = \arg\max_k [Q_t(k) + U_t(k)], \tag{1}$$

where a_t is the action chosen at time t, $Q_t(k)$ is the expected reward of action k at time t, and $U_t(k)$ is the upper confidence bound of the reward that plays the role of an uncertainty bonus. In a Bayesian variant of UCB (Srinivas et al., 2010), $Q_t(k)$ corresponds to the posterior mean and $U_t(k)$ is proportional to the posterior standard deviation $\sigma_t(k)$. Returning to the restaurant example, even if both restaurants have the same expected value ($Q_t(\text{new}) = Q_t(\text{old})$), UCB would initially prefer the new one since it has greater uncertainty ($U_t(\text{new}) > U_t(\text{old})$).

Random exploration strategies introduce randomness into choice behavior, causing the agent to sometimes explore less favorable options. For example, when you move to a new neighborhood, you might initially pick restaurants at random until you learn which ones are good. Random exploration is thus sensitive to the *total uncertainty* of the available options, increasing choice stochasticity when option values are more uncertain. This can cause choice variability to track payoff variability, a phenomenon sometimes referred to as the payoff variability effect (Myers and Sadler, 1960; Busemeyer and Townsend, 1993; Erev and Barron, 2005).

One prominent instantiation of random exploration in reinforcement learning is Thompson sampling (Thompson, 1933), which samples values randomly from the posterior value distribution of each action and then chooses greedily with respect to the sampled values:

$$\widetilde{Q}_t(k) \sim p(Q_t(k)) \tag{2}$$

$$a_t = \arg\max_k \widetilde{Q}_t(k), \tag{3}$$

where $p(.)$ is the posterior value distribution and $\widetilde{Q}_t(k)$ is the sampled value for arm k at time t. Returning to the

neighborhood example, the familiar restaurants in your old neighborhood have narrow value distributions around their expected values (low total uncertainty). This will cause Thompson sampling to consistently draw samples $\widetilde{Q}_t(k)$ that are close to their expected values, which will often result in choosing the same restaurant, namely the one with the highest expected value. In contrast, the unfamiliar restaurants in the new neighborhood have wide value distributions (high total uncertainty), which will result in significant variation in the Thompson samples $\widetilde{Q}_t(k)$ and a corresponding variation in the chosen restaurant.

Directed and random exploration strategies confer separate computational advantages, which has led researchers in reinforcement learning to develop algorithms that use a hybrid of UCB and Thompson sampling (Chapelle and Li, 2011; May et al., 2012). Correspondingly, recent evidence suggests that people also employ a combination of directed and random exploration (Wilson et al., 2014; Gershman, 2018a). A study by Gershman (2018b) used a two-armed bandit task to show that human choices are consistent with a particular hybrid of UCB and Thompson sampling. Furthermore, the study showed that different uncertainty computations underlie each exploration strategy, with the relative uncertainty between the two options driving directed exploration, and the total uncertainty of the two options driving random exploration. This led us to hypothesize the existence of dissociable neural implementations of both strategies in the brain. At least three lines of evidence support this claim. First, dopamine genes with anatomically distinct expression profiles are differentially associated with directed and random exploration (Gershman and Tzovaras, 2018). Second, transcranial magnetic stimulation of right rostrolateral prefrontal cortex (RLPFC) affects directed, but not random, exploration (Zajkowski et al., 2017). Third, directed and random exploration have different developmental trajectories (Somerville et al., 2017).

In the present study, we used functional MRI to probe the neural underpinnings of the uncertainty computations that influence directed and random exploration. Subjects performed a two-armed bandit task in which each arm was either “safe”, meaning it delivered the same reward during the whole block, or “risky”, meaning it delivered variable rewards. This allowed us to separate the effects of relative and total uncertainty and examine how their neural correlates influence directed and random exploration. We found that relative uncertainty is encoded in right RLPFC and total uncertainty is encoded in left anterior insula. The neural signal in right RLPFC predicted cross-trial variability in

directed but not random exploration, whereas the neural signal in left anterior insula predicted cross-trial variability in random but not directed exploration. Furthermore, cross-subject differences in the encoding of total uncertainty predicted differences in the extent to which subjects employ random exploration. These results build on previous findings that implicate right RLPFC in directed exploration (Badre et al., 2012; Zajkowski et al., 2017) and point to left anterior insula as an anatomical locus of random exploration. By linking activity in both regions with human choices via a hybrid model, our work provides new insight into the distinct uncertainty computations performed by the brain and their role in guiding behavior.

Results

Relative and total uncertainty guide directed and random exploration

We scanned 31 human subjects (17 female, ages 18-35) using functional MRI while they performed a two-armed bandit task in which subjects are informed about the riskiness of each option (Gershman, 2018b). On each trial, subjects saw the labels of the two options, each of which could be either "safe" (S) or "risky" (R; Figure 1A). A safe option always delivered the same reward for the duration of the block, while a risky option delivered Gaussian-distributed rewards around a mean that remained fixed for the duration of the block (Figure 1B). We denote the trial types by the pair of option labels (e.g., on "RS" trials, the left option is risky and the right one is safe). Trial types and average rewards varied randomly across blocks. Subjects were explicitly informed of the statistics of the task and performed four practice blocks before entering the scanner.

Importantly, this task design allowed us to independently measure the effect of different types of uncertainty on subject choices. Directed and random exploration predict different effects for these manipulations, which can be illustrated by considering the probability of choosing the left option, P(choose left), as a function of the relative expected value, V = Q(left) - Q(right) (the choice function; Figure 1C and D). Note that our decision to model the left option first is arbitrary and without loss of generality, since the two options are symmetric (P(choose left) = 1 - P(choose right)) and the same analysis would apply if we reverse their order.

RS and SR trials manipulate relative uncertainty (greater for the left option on RS trials and greater for the right option on SR trials) while controlling for total uncertainty (identical across RS and SR trials). A strategy that is insensitive to relative uncertainty such as softmax exploration would be indifferent between the two options (P(choose left) = 0.5) when they have equal values (V = 0). In contrast, UCB predicts a bias towards the risky option, preferring to choose left on RS trials and right on SR trials, even when the relative exptected value might dictate otherwise. This would manifest as an opposite intercept shift in the choice probability function of RS and SR trials, such that P(choose left) > 0.5 when V = 0 on RS trials and P(choose left) < 0.5 when V = 0 on SR trials (Figure 1C).

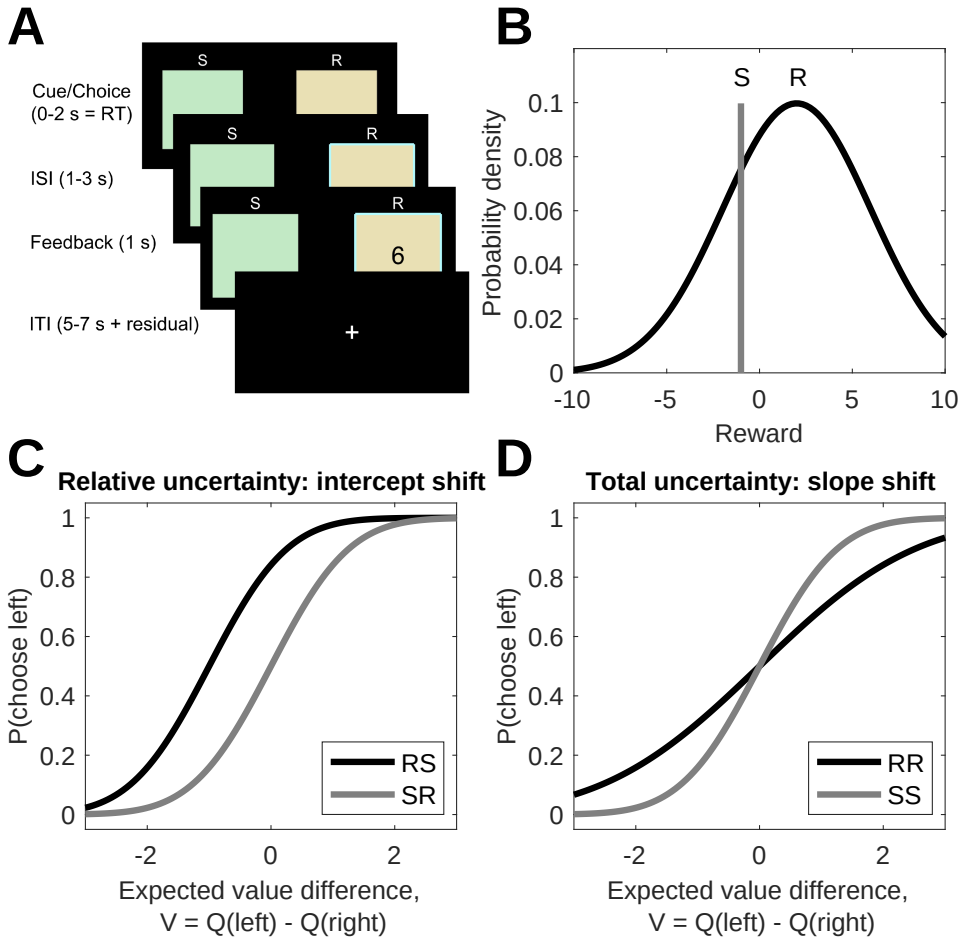

Figure 1. **Experimental Design and Predictions.**

(A) Trial structure. Subjects choose between two options, each labeled as either safe (S) or risky (R). After they make a choice, they receive feedback in the form of points. Option labels remain constant during the block.

(B) Reward structure. Risky options deliver rewards drawn from a Gaussian distribution whose mean remains constant during the block. Safe options deliver the same reward during the block. The means of both options are resampled from the zero-mean Gaussian at the start of each block.

(C) Directed exploration (UCB) predicts a bias towards the uncertain option, which shifts the choice probability function in the opposite directions for RS and SR trials.

(D) Random exploration (Thompson sampling) predicts more randomness when uncertainty is high, which reduces the slope of the choice probability function for RR compared to SS trials.

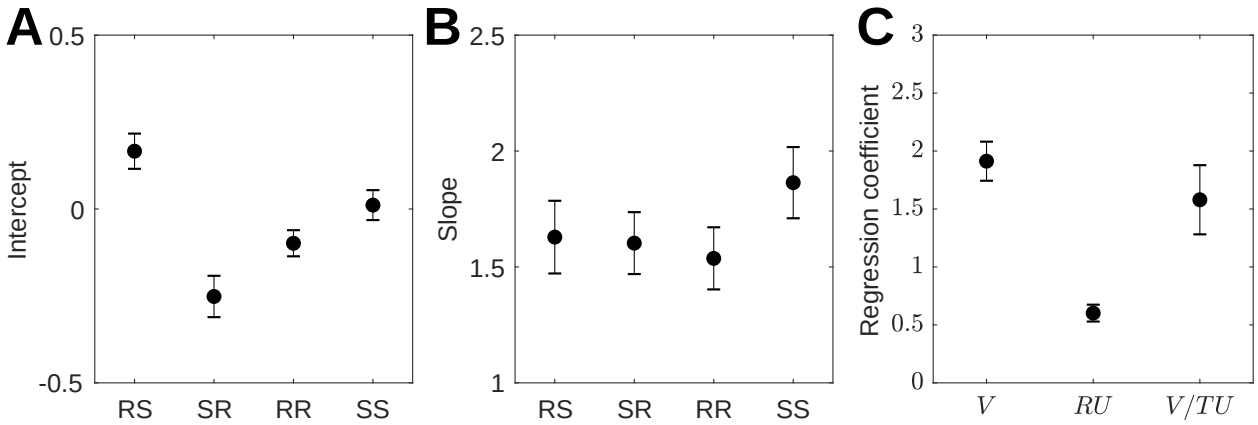

Figure 2. **Probit regression results.**

(A) Intercept and (B) slope of choice probability function fit for each condition using maximum likelihood estimation. (C) Maximum likelihood estimates of coefficients w_1, w_2, w_3 for each regressor in the hybrid model (Eq. 4). Error bars are within-subject standard errors. V = difference between the two estimated option values; RU = relative uncertainty between the two value estimates, TU = total uncertainty of the two value estimates.

In contrast, RR and SS trials manipulate total uncertainty (high on RR trials and low on SS trials) while controlling for relative uncertainty (identical across RR and SS trials). A strategy that is insensitive to total uncertainty such as UCB would predict the same choice function for both trial types. In contrast, Thompson sampling predicts more stochastic choices when there is more uncertainty, resulting in more random choices (P(choose left) closer to 0.5) even when the relative expected value strongly favors one option (V far from 0). This would manifest as a shallower slope of the choice probability function of RR compared to SS trials (Figure 1D).

Consistent with the theory, we found that manipulating relative uncertainty (RS vs. SR) shifted the intercept of the choice probability function (Figure 2A): the intercept for RS trials was significantly greater than the intercept for SR trials ($F(1,9711) = 21.0, p = 0.000005$). Moreover, the intercept for RS trials was significantly greater than 0 ($F(1,9711) = 10.8, p = 0.001$), while the intercept for SR trials was significantly less than 0 ($F(1,9711) = 17.9, p = 0.00002$). There was only a small effect of total uncertainty on the intercept (RR vs. SS, $F(1,9711) = 4.1, p = 0.04$). This indicates that, regardless of relative expected value, subjects showed a bias towards the risky option, consistent with UCB.

Conversely, manipulating total uncertainty (RR vs. SS) altered the slope of the choice probability function (Figure 2B): the slope for RR trials is smaller than the slope for SS trials ($F(1,9711) = 3.4, p = 0.07$). While the effect is small due to the small sample size, it is in the right direction and is consistent with previous replications of this experiment (Gershman, 2018b; Gershman and Tzovaras, 2018). There was no effect of relative uncertainty (RS vs. SR) on the slope ($F(1,9711) = 0.06, p = 0.8$). This indicates that when both options were risky, subjects were less sensitive to their relative reward advantage, consistent with Thompson sampling.

To examine how relative and total uncertainty influence directed and random exploration on a trial-by-trial basis, we modeled subject choices using a probit regression model (Gershman, 2018b):

$$P(a_t = \text{left}|\mathbf{w}) = \Phi(w_1 \text{V}_t + w_2 \text{RU}_t + w_3 \text{V}_t/\text{TU}_t), \tag{4}$$

where $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function and the regressors are:

- Estimated value difference, $\mathrm{V}_t = Q_t(\text{left}) - Q_t(\text{right})$.
- Relative uncertainty, $\mathrm{RU}_t = \sigma_t(\text{left}) - \sigma_t(\text{right})$.
- Total uncertainty, $\mathrm{TU}_t = \sqrt{\sigma_t^2(\text{left}) + \sigma_t^2(\text{right})}$.

Here $Q_t(k)$ corresponds to the posterior expected value of option k and $\sigma_t(k)$ is the posterior standard deviation around that expectation, proportional to the uncertainty bonus in UCB.

Gershman (2018a) showed that this corresponds to a hybrid of UCB and Thompson sampling that reduces to pure UCB when $w_3 = 0$ and to pure Thompson sampling when $w_2 = 0$. If subjects use both strategies, the model predicts that all three regressors will have a significant effect on choices ($w_1 > 0, w_2 > 0, w_3 > 0$). Correspondingly, the maximum likelihood estimates of all three coefficients were significantly greater than 0 (Figure 2C, $p < 10^{-6}$ for each coefficient). Taken together, these results replicate previous findings (Gershman, 2018b) and show that people employ both directed and random exploration, driven by relative and total uncertainty, respectively.

Neural correlates of relative and total uncertainty

Next, we asked whether relative and total uncertainty are computed in distinct anatomical loci. We performed an unbiased whole-brain univariate analysis using a general linear model (GLM) with model-derived trial-by-trial estimates of absolute relative uncertainty ($|\mathrm{RU}_t|$), total uncertainty (TU_t), and trial ordinal within the block (trial = t) as non-orthogonalized impulse regressors at trial onset (see Experimental Procedures). Note that the uncertainty of both options gradually decreases over the course of each block as more values are sampled for each option, which causes both $|\mathrm{RU}_t|$ and TU_t to decrease monotonically. To account for this monotonic decrease, we computed the group-level contrasts between $|\mathrm{RU}_t|$ and t (RU - trial) and between TU_t and t (TU - trial). We report whole-brain t-maps after thresholding single voxels at $p < 0.001$ (uncorrected) and thresholding clusters at extent >= 100 voxels. We chose to use an uncorrected threshold in order to identify candidate regions for our subsequent confirmatory

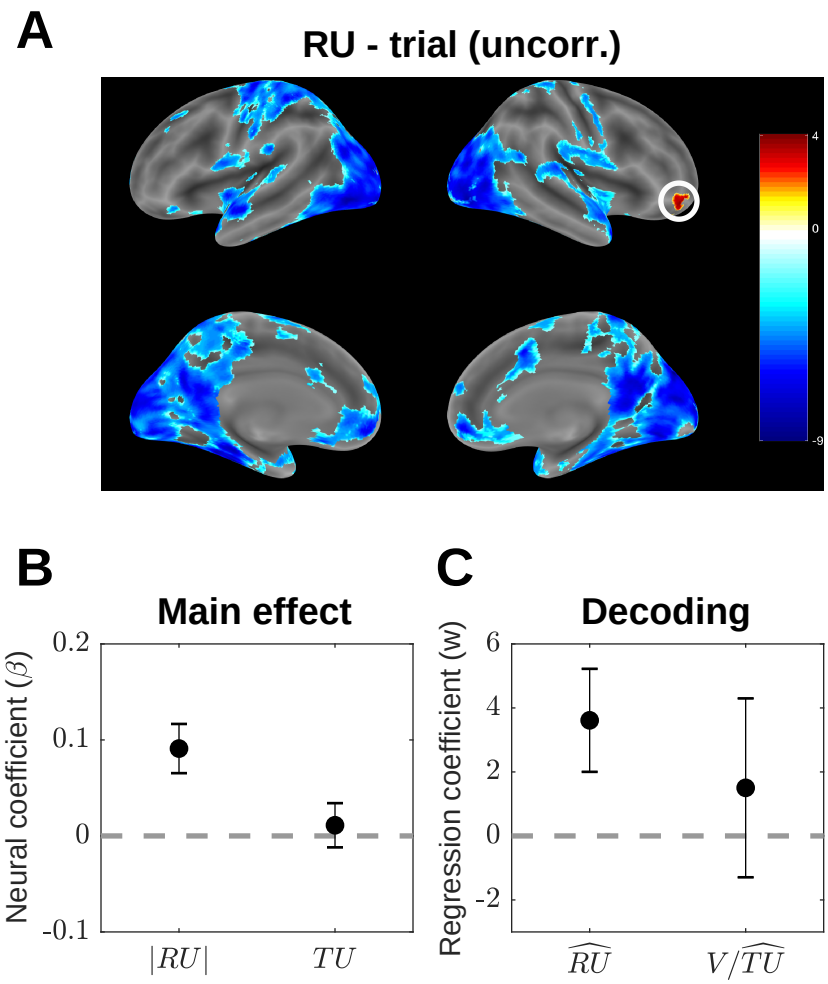

Figure 3. **Right RLPFC encodes RU but not TU.**

(A) Whole-brain contrast between RU and trial number. Single voxels were thresholded at $p < 0.001$ (uncorrected) and clusters were thresholded at extent >= 100 voxels. The color scale represents t-values across subjects. The circled ROI in right RLPFC (MNI 34 50 -10) was used in the subsequent analyses (10 mm sphere around the peak voxel).

(B) Neural regression coefficients (betas) from the GLM for the parametric modulators $|\mathrm{RU}_t|$ ($\beta_{|\mathrm{RU}|}$) and TU_t (β_{TU}) at trial onset, averaged across voxels in the ROI and across subjects. RU = relative uncertainty; TU = total uncertainty.

(C) Regression coefficients (w_4 in Eq. 11 and w_5 in Eq. 12, respectively) from the augmented models with the neurally-decoded relative uncertainty ($\widehat{\mathrm{RU}}$) or total uncertainty ($\widehat{\mathrm{TU}}$). V = value difference between the two option estimates. Error bars are cross-subject standard errors.

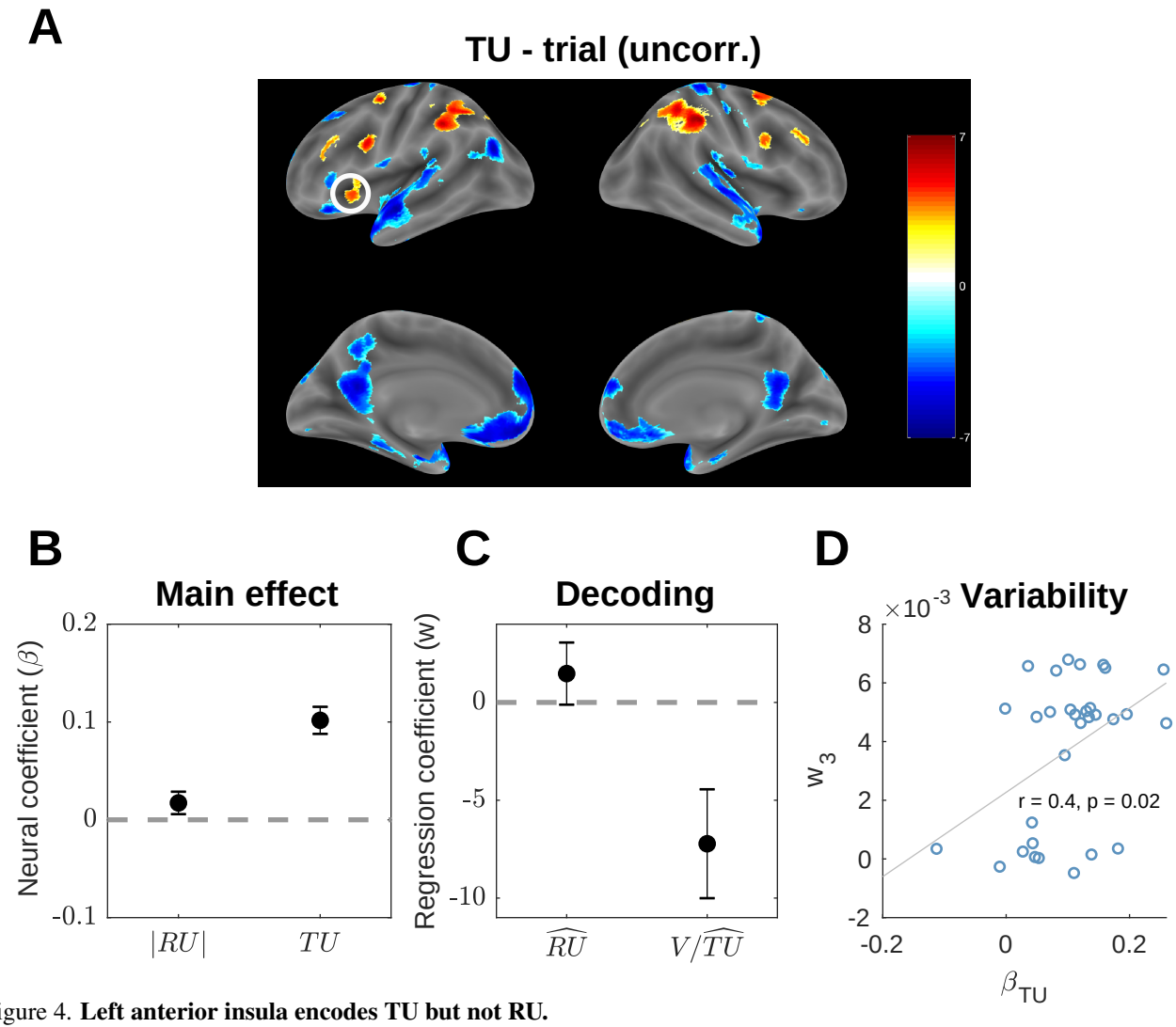

Figure 4. **Left anterior insula encodes TU but not RU.**

(A) Whole-brain contrast between TU and trial. Single voxels were thresholded at $p < 0.001$ (uncorrected) and clusters were thresholded at extent >= 100 voxels. The color scale represents t-values across subjects. The circled ROI in left anterior insula (MNI -30 22 2) was used in subsequent analyses (10 mm sphere around the peak voxel).

(B) and (C) same as in Figure 3 but using the left anterior insula.

(D) Between-subject correlation between the neural coefficient for TU (β_{TU}) and the extent to which the subject employs random exploration (w_3 in Eq. 4). r, Pearson's correlation coefficient.

analyses (corrected whole-brain results are reported in the Supplemental Information).

For the RU - trial contrast, we did not find any positive regions that survived FWE correction (Figure S1A). We therefore used the *a priori* region of interest (ROI) from Badre et al. (2012) that was found to track RU (right RLPFC, MNI 36 56 -8; Figure S2A). We found a significant effect of RU ($t(30) = 3.1, p = 0.005$) but not TU ($t(30) = 0.10, p = 0.91$) in the right RLPFC ROI, thus replicating the previous findings (Figure S2B).

We then considered the uncorrected RU - trial results with minimum cluster extent >= 100 voxels (Figure 3A and Table S2) and found a similar cluster in right RLPFC (MNI 34 50 -10) which showed a significant effect for RU ($t(30) = 3.6, p = 0.001$) but not TU ($t(30) = 0.48, p = 0.63$; Figure 3B). We therefore restrict our subsequent analyses of relative uncertainty to that ROI.

For the TU - trial contrast (Figure S1B and Table S3), we found bilateral positive clusters in inferior parietal gyrus and primary motor cortex, in addition to unilateral clusters in left anterior insula and right superior premotor cortex. Importantly, neural activity in these regions did not track RU (Figure 4B, $|t(30)| < 1.5, p > 0.14$ for all ROIs).

We also used the *a priori* ROI from Badre et al. (2012) that was found to track TU (right DLPFC, MNI 40 30 34; Figure S3A). Similarly to their results, we also found that right DLPFC tracks both RU ($t(30) = 2.24, p = 0.03$) and TU ($t(30) = 4.32, p = 0.0002$). Since we did not find right DLPFC in our TU - trial contrast and the main effect in right DLPFC was not specific to TU, we used all ROIs from our contrast in subsequent analyses of total uncertainty. To remain consistent with the relative uncertainty analysis, we consider ROIs from the uncorrected TU - trial contrast with minimum cluster extent >= 100 voxels (Figure 4A and Table S3).

For a stronger confirmation that relative and total uncertainty correlate with different brain networks, we also inspected the contrasts RU - TU and TU - RU (Figure S1C and D). This revealed regions that largely overlapped with the regions identified by the RU - trial and TU - trial contrasts. Nevertheless, we chose not to use ROIs from these contrasts as they confound negative and positive correlations.

Subjective estimates of relative and total uncertainty drive choices

If the ROIs identified by the GLM encode relative and total uncertainty, then we should be able to use their activations to decode trial-by-trial subjective estimates of RU and TU. In particular, on any given trial, a subject's estimate of RU and TU might differ from the ideal observer estimates used in the hybrid model (Eq. 4). This could occur for a number of reasons, such as neural noise, inattention, or a suboptimal learning rate. Importantly, any deviation from the ideal observer estimates would result in a corresponding deviation of the subject's choices from the model predictions. Therefore if we augment the hybrid model to include the neurally-decoded subjective estimates of RU and TU (denoted by $\widehat{\mathrm{RU}}$ and $\widehat{\mathrm{TU}}$, respectively), then we should arrive at more accurate predictions of the subject's choices (see Experimental Procedures).

Indeed, this is what we found: including the decoded $\widehat{\mathrm{RU}}$ (Eq. 11) from right RLPFC significantly improved the trial-by-trial predictions of subject choices. This was true both when we used our right RLPFC ROI (MNI 34 50 -10; Table S4 and Figure 3C; $p = 0.02$, likelihood ratio test) and the right RLPFC ROI from Badre et al. (2012) (MNI 36 56 -8; Figure S2C; $p = 0.003$, likelihood ratio test). Importantly, decoding $\widehat{\mathrm{TU}}$ from either ROI and augmenting the model with $V/\widehat{\mathrm{TU}}$ (Eq. 12) did not improve choice predictions ($p > 0.1$ for both ROIs).

Similarly, we found a significant effect when including $V/\widehat{\mathrm{TU}}$ (Eq. 12), with $\widehat{\mathrm{TU}}$ decoded from either left primary motor cortex (MNI 22 -2 52; $p = 0.03$), right premotor cortex (MNI 46 8 30; $p = 0.00009$), or left anterior insula (MNI -30 22 2; Table S5 and Figure 4C; $p = 0.0002$, likelihood ratio test, all p-values Bonferroni corrected for 9 comparisons across all positive ROIs). However, decoding $\widehat{\mathrm{TU}}$ from right DLPFC (MNI 40 30 34) did not improve choice predictions (Figure S3C; $p = 0.4$). Importantly, decoding $\widehat{\mathrm{RU}}$ from our ROIs did not improve choice predictions ($p > 0.05$ for all 3 ROIs). Together, these results show that variability in the neural representations of uncertainty in the corresponding regions affects choices, suggesting that those representations are used in a downstream decision computation.

Cross-subject differences in neural encoding of uncertainty predict differences in exploration strategy

Finally, we sought to characterize the relationship between inter-subject variability in exploration strategy and neural activity. We hypothesized that the strength of neural encoding of RU and TU will predict the extent to which subjects engage in directed and random exploration, respectively. For example, if the neural signal for TU is noisier or weaker for a given subject, then we would expect their choices to have a smaller random exploration component. The former corresponds to the beta coefficient for TU derived from the neural GLM (β_{TU}), while the latter corresponds to the regression coefficient of V/TU in the hybrid model (w_3 in Eq. 4). Across subjects, we thus correlated w_2 with $\beta_{|\mathrm{RU}|}$ for right RLPFC and w_3 with β_{TU} for left primary motor cortex, right premotor cortex, and left anterior insula.

We found no relationship for RU in our right RLPFC ROI (MNI 34 50 -10; $r = -0.09, p = 0.61$) or in the right RLPFC ROI from Badre et al. (2012) (MNI 36 56 -8; $r = -0.16, p = 0.4$, Pearson correlation across 31 subjects). We found a significant relationship for TU in left anterior insula (MNI -30 22 2; $r = 0.43, p = 0.047$, Pearson correlation across 31 subjects, Bonferroni corrected for 3 comparisons) but not the other two ROIs (Table S6 and Figure 4D). Importantly, there was no relationship for TU in right RLPFC ($r = 0.02, p = 0.9$). These results point to left anterior insula as the neural locus of total uncertainty, which drives random exploration in a subject-specific manner.

Finally, we confirmed that including $\widehat{\mathrm{RU}}$ decoded from right RLPFC (MNI 34 50 -10) and $V/\widehat{\mathrm{TU}}$ decoded from left anterior insula (MNI -30 22 2) in the same model (Eq. 13) improved choice predictions beyond the improvement from including either estimate alone ($LRStat = 18.1, p = 0.00002$ when comparing with $\widehat{\mathrm{RU}}$ alone, Eq. 11; and $LRStat = 5.3, p = 0.02$ when comparing with $V/\widehat{\mathrm{TU}}$ alone, Eq. 12, likelihood ratio test).

Discussion

Balancing exploration and exploitation lies at the heart of decision making, and understanding the neural circuitry that underlies different forms of exploration is central to understanding how the brain makes choices in the real world. Here we show that human choices are consistent with a particular hybrid of directed and random exploration strategies,

driven respectively by the relative and total uncertainty of the options. This dissociation between the two uncertainty computations predicted by the model was reified as an anatomical dissociation between their neural correlates. Our GLM results confirm the previously identified role of right RLPFC in encoding relative uncertainty (Badre et al., 2012) and influencing directed exploration (Zajkowski et al., 2017), while our decoding analysis further suggests that the subjective estimate of relative uncertainty in RLPFC is used in the downstream decision computation. Our results also point to a previously unidentified role of left anterior insula in encoding the subjective estimate of total uncertainty in a way that both influences choices and accounts for inter-subject variability in exploration strategy.

Previous studies have also found a signature of exploration in RLPFC, also referred to as frontopolar cortex (Daw et al., 2006; Boorman et al., 2009; Badre et al., 2012; Beharelle et al., 2015), however, with the exception of Badre et al. (2012), these studies did not disentangle different exploration strategies or examine their relation to uncertainty. More pertinent to our study, Zajkowski et al. (2017) reported that inhibiting right RLPFC reduces directed but not random exploration. This is consistent with our finding that right RLPFC encodes a subjective estimate of relative uncertainty which underlies the directed-exploration component of choices in our model. Disrupting activity in right RLPFC can thus be understood as introducing noise into or reducing the subjective estimate of relative uncertainty, resulting in choices that are less consistent with directed exploration.

Our finding that total uncertainty is encoded in left anterior insula is somewhat puzzling, given that this region is traditionally associated with interoception and affective processing. A related finding was reported by Blanchard and Gershman (2018) who used a task that separates exploration and exploitation; that study found a neural correlate of pure exploration in bilateral anterior insula and dorsal anterior cingulate cortex, both part of a broader salience network (Menon and Uddin, 2010). Another study by Ohira et al. (2013) found that activity in left anterior insula was correlated with blood epinephrine level as well as choice stochasticity, which suggests a link between uncertainty, arousal, and random exploration. Our results build on those findings by establishing a specific role of left anterior insula in driving the random-exploration component of choices by computing a subjective estimate of total uncertainty.

A notable feature of our results is that the fidelity of the total uncertainty signal in anterior insula predicts the subject's

reliance on random exploration, but no such relationship was found between relative uncertainty in RLPFC and directed exploration. This contrasts with the finding of Badre et al. (2012) that explorers tend to have a stronger representation of relative uncertainty in RLPFC than non-explorers. Furthermore, Badre et al. (2012) do not report correlates of total uncertainty (mean uncertainty in their study) in left anterior insula. These discrepancies could possibly arise from the fact that in their study, relative and total uncertainty were not manipulated independently, and only relative uncertainty was used to model choices and to distinguish between explorers and non-explorers.

The neural and behavioral dissociation between relative and total uncertainty found in our study points to potential avenues for future research that could establish a double dissociation between the corresponding brain regions. Temporarily disrupting activity in right RLPFC should affect directed exploration (reducing w_2 in Eq. 4), while leaving random exploration intact (not changing w_3 in Eq. 4). Conversely, disrupting left anterior insula should affect random but not directed exploration. This would expand upon the RLPFC results of Zajkowski et al. (2017) by establishing a causal role for both regions in the corresponding uncertainty computations and exploration strategies.

In summary, we show that humans tackle the exploration-exploitation trade-off using a combination of directed and random exploration strategies driven by neurally-dissociable uncertainty computations. Relative uncertainty was encoded in right RLPFC and influenced directed exploration, while total uncertainty was encoded in left anterior insula and influenced random exploration. Subjective trial-by-trial estimates decoded from both regions predicted subject responding, while cross-subject differences in the anterior insula signal predicted differences in reliance on random exploration. Together, these results suggest that the signals from both regions are integrated by downstream circuits to ultimately compute choice, thus providing an important step towards understanding how the brain implements uncertainty-driven exploration.

Experimental Procedures

Subjects

We recruited 31 subjects (17 female) from the Cambridge community. All subjects were healthy, ages 18-35, right-handed, with normal or corrected vision, and no neuropsychiatric pre-conditions. Subjects were paid $50.00 for their participation plus a bonus based on their performance. The bonus was the number of points from a random trial paid in dollars (negative points were rounded up to 1). All subjects received written consent and the study was approved by the Harvard Institutional Review Board.

Experimental Design and Statistical Analysis

We used the two-armed bandit task described in Gershman (2018b). On each block, subjects played a new pair of bandits for 10 trials. Each subject played 32 blocks, with 4 blocks in each scanner run (for a total of 8 runs per subject). On each trial, subjects chose an arm and received reward feedback (points delivered by the chosen arm). Subjects were told to pick the better arm in each trial. To incentivize good performance, subjects were told that at the end of the experiment, a trial will be drawn randomly and they will receive the number of points in dollars (negative rewards were rounded up to 1).

On each block, the mean reward $\mu(k)$ for each arm k was drawn randomly from a Gaussian with mean 0 and variance $\tau_0^2(k) = 100$. Arms on each block were designated as "risky" (R) or "safe" (S), with all four block conditions (RS, SR, RR, and SS) counterbalanced and randomly shuffled within each run (i.e. each run had one of each block condition in a random order). A safe arm delivered the same reward $\mu(S)$ on each trial during the block. A risky arm delivered rewards sampled randomly from a Gaussian with mean $\mu(R)$ and variance $\tau^2(R) = 16$. The type of each arm was indicated to subjects by the letter R or S above the corresponding box. In order to make it easier for subjects to distinguish between the arms within and across blocks, each box was filled with a random color that remained constant throughout the blocks and changed between blocks. The color was not informative of rewards.

Subjects were given the following written instructions.

In this task, you have a choice between two slot machines, represented by colored boxes. When you choose one of the slot machines, you will win or lose points. One slot machine is always better than the other, but choosing the same slot machine will not always give you the same points. ***Your goal is to choose the slot machine that you think will give you the most points****. Sometimes the machines are "safe" (always delivering the same points), and sometimes the machines are "risky" (delivering variable points). Before you make a choice, you will get information about each machine: "S" indicates SAFE, "R" indicates RISKY. The "safe" or "risky" status does not tell you how rewarding a machine is. A risky machine could deliver more or less points on average than a safe machine. You cannot predict how good a machine is simply based on whether it is considered safe or risky. Some boxes will deliver negative points. In those situations, you should select the one that is least negative. In the MRI scanner, you will play 32 games, each with a different pair of slot machines. Each game will consist of 10 trials. Before we begin the actual experiment, you will play a few practice games. Choose the left slot machine by pressing with your index finger and the right slot machine by pressing with your middle finger. You will have 2 seconds to make a choice. To encourage you to do your best, at the end of the MRI experiment, a random trial will be chosen and you will be paid the number of points you won on that trial in dollars.* ***You want to do as best as possible every time you make a choice!***

The event sequence within a trial is shown in Figure 1A. At trial onset, subjects saw two boxes representing the two arms and chose one of them by pressing with their index finger or their middle finger on a response button box. The chosen arm was highlighted and, after a random inter-stimulus interval (ISI), they received reward feedback as the number of points they earned from that arm. No feedback was provided for the unchosen arm. Feedback remained on the screen for 1 s, followed by a random inter-trial interval (ITI) and the next trial. A white fixation cross was displayed during the ITI. If subjects failed to respond within 2 s, they were not given feedback and after the ISI, they directly entered the ITI with a red fixation cross. Otherwise, the residual difference between 2 s and their reaction time was added to the following ITI. Each block was preceded by a 6-second inter-block-interval, during which subjects saw a sign, "New game is starting," for 3 s, followed by a 3-second fixation cross. A 10-second fixation cross was

added to the beginning and end of each run to allow for scanner stabilization and hemodynamic lag, respectively. ISIs and ITIs were pregenerated by drawing uniformly from the ranges 1-3 s and 5-7 s, respectively. Additionally, ISIs and ITIs in each run were uniformly scaled such that the total run duration is exactly 484 s, which is the expected run length assuming an average ISI (2 s) and an average ITI (6 s). This accounted for small deviations from the expected run duration and allowed us to acquire 242 whole-brain volumes during each run (TR = 2 s). The experiment was implemented using the PsychoPy toolbox (Peirce, 2007).

Belief updating model

Following Gershman (2018b), we assumed subjects approximate an ideal Bayesian observer that tracks the expected value and uncertainty for each arm. Since rewards in our task are Gaussian-distributed, these correspond to the posterior mean $Q_t(k)$ and variance $\sigma_t^2(k)$ of each arm k, which can be updated recursively on each trial t using the Kalman filtering equations:

$$Q_{t+1}(a_t) = Q_t(a_t) + \alpha_t[r_t - Q_t(a_t)] \tag{5}$$

$$\sigma_{t+1}^2(a_t) = \sigma_t^2(a_t) - \alpha_t \sigma_t^2(a_t), \tag{6}$$

where a_t is the chosen arm, r_t is the received reward, and the learning rate α_t is given by:

$$\alpha_t = \frac{\sigma_t^2(a_t)}{\sigma_t^2(a_t) + \tau^2(a_t)}. \tag{7}$$

We initialized the values with the prior means, $Q_t(k) = 0$ for all k, and variances with the prior variances, $\sigma_1^2(k) = \tau_0^2(k)$. Subjects were informed of those priors and performed 4 practice blocks (40 trials) before entering the scanner to familiarize themselves with the task structure. Kalman filtering is the Bayes-optimal algorithm for updating the values and uncertainties given the task structure and has been previously shown to account well for human choices in bandit tasks (Daw et al., 2006; Schulz et al., 2015; Speekenbrink and Konstantinidis, 2015; Gershman, 2018a).

Choice probability analysis

We fit the coefficients $\mathbf{w}$ of the hybrid model (Eq. 4) using mixed-effects maximum likelihood estimation (fitglme in MATLAB) using all non-timeout trials.

In addition to the hybrid model, we also fit a model of choices as a function of experimental condition to obtain the slope and intercept of the choice probability function:

$$P(a_t = \text{left}|\mathbf{w}) = \Phi\left(\sum_j w_1^j \pi_{tj} + w_2^j \pi_{tj} V_t\right), \tag{8}$$

where j is the experimental condition (RS, SR, RR, or SS), and $\pi_{tj} = 1$ if trial t is assigned to condition j, and 0 otherwise. We plotted the w_1 terms as the intercepts and the w_2 terms as the slopes.

fMRI Data Acquisition

We followed the same protocol as described previously (Tomov et al., 2018). Scanning was carried out on a 3T Siemens Magnetom Prisma MRI scanner with the vendor 32-channel head coil (Siemens Healthcare, Erlangen, Germany) at the Harvard University Center for Brain Science Neuroimaging. A T1-weighted high-resolution multi-echo magnetization-prepared rapid-acquisition gradient echo (ME-MPRAGE) anatomical scan (van der Kouwe et al., 2008) of the whole brain was acquired for each subject prior to any functional scanning (176 sagittal slices, voxel size = 1.0 x 1.0 x 1.0 mm, TR = 2530 ms, TE = 1.69 - 7.27 ms, TI = 1100 ms, flip angle = 7°, FOV = 256 mm). Functional images were acquired using a T2*-weighted echo-planar imaging (EPI) pulse sequence that employed multiband RF pulses and Simultaneous Multi-Slice (SMS) acquisition (Moeller et al., 2010; Feinberg et al., 2010; Xu et al., 2013). In total, 8 functional runs were collected for each subject, with each run corresponding to 4 task blocks, one in each condition (84 interleaved axial-oblique slices per whole brain volume, voxel size = 1.5 x 1.5 x 1.5 mm, TR = 2000 ms, TE = 30 ms, flip angle = 80°, in-plane acceleration (GRAPPA) factor = 2, multi-band acceleration factor = 3, FOV = 204 mm). The initial 5 TRs (10 s) were discarded as the scanner stabilized. Functional slices were oriented to a 25 degree tilt towards coronal from AC-PC alignment. The SMS-EPI acquisitions used the CMRR-MB pulse sequence from the University of Minnesota.

All 31 scanned subjects were included in the analysis. We excluded runs with excessive motion (greater than 2 mm translational motion or greater than 2 degrees rotational motion). Four subjects had a single excluded run and two

additional subjects had two excluded runs.

fMRI Preprocessing

As in our previous work (Tomov et al., 2018), functional images were preprocessed and analyzed using SPM12 (Wellcome Department of Imaging Neuroscience, London, UK). Each functional scan was realigned to correct for small movements between scans, producing an aligned set of images and a mean image for each subject. The high-resolution T1-weighted ME-MPRAGE images were then co-registered to the mean realigned images and the gray matter was segmented out and normalized to the gray matter of a standard Montreal Neurological Institute (MNI) reference brain. The functional images were then normalized to the MNI template (resampled voxel size 2 mm isotropic), spatially smoothed with a 8 mm full-width at half-maximum (FWHM) Gaussian kernel, high-pass filtered at 1/128 Hz, and corrected for temporal autocorrelations using a first-order autoregressive model.

Univariate Analysis

Our hypothesis was that different brain regions perform the two kinds of uncertainty computations (relative and total uncertainty), which in turn drive the two corresponding exploration strategies (directed exploration, operationalized as UCB, and random exploration, operationalized as Thompson sampling). We therefore defined a general linear model (GLM, Table S1) with absolute relative uncertainty, $|\mathrm{RU}_t|$, and total uncertainty, TU_t, as parametric modulators for an impulse regressor at trial onset (trial_onset). Following (Badre et al., 2012), we used $|\mathrm{RU}_t|$ instead of RU_t to account for our arbitrary choice of arm 1 and arm 2 (note that TU_t is always positive). The trial_onset regressor was also parametrically modulated by the trial number within the block ($t = 1..10$), since both $|\mathrm{RU}_t|$ and TU_t tend to decrease monotonically during the block, which would be confounded with any activity that tracks time within the block. This regressor was only included on trials on which the subject responded within 2 s of trial onset. We included a separate regressor at trial onset for trials on which the subject timed out that was not parametrically modulated (trial_onset_timeout), since failure to respond could be indicative of failure to perform the necessary uncertainty computations. In order to control for any motor-related activity that might be captured by those regressors due to the hemodynamic lag, we also included a separate trial onset regressor on trials on which the subject chose arm 1

(trial_onset_chose_1). Finally, to account for response-related activity and feedback-related activity, we included regressors at reaction time (button_press) and feedback onset (feedback_onset), respectively. All regressors were impulse regressors (duration = 0 s) convolved with the canonical hemodynamic response function (HRF). The parametric modulators were not orthogonalized (Mumford et al., 2015). As is standard in SPM, there was a separate version of each regressor for every scanner run. Additionally, there were six motion regressors and an intercept regressor.

In order to control for confounds of time within block, we followed our previous work (Tomov et al., 2018) and computed group-level contrasts between the uncertainty regressors and trial number (RU - trial and TU - trial). An alternative approach would be to orthogonalize the uncertainty regressors with respect to trial number and compute the single-regressor contrasts RU and TU. However, previous work (Mumford et al., 2015) suggests that this would not make a difference for the beta coefficients of RU and TU, and hence we preferred the contrast approach.

For group-level whole-brain analyses, we performed t-contrasts with single voxels thresholded at $p < 0.001$ and cluster family-wise error (FWE) correction applied at $\alpha = 0.05$. Since this did not reveal any clusters for RU, for ROI selection in subsequent analyses, we used uncorrected results (extent >= 100 voxels) for both RU and TU. We labeled clusters based on the peak voxel labels from the deterministic Automated Anatomical Labeling (AAL2) atlas (Tzourio-Mazoyer et al., 2002; Rolls et al., 2015). For clusters whose peak voxels were not labeled successfully by AAL2, we consulted the SPM Anatomy Toolbox (Eickhoff et al., 2005) and the CMA Harvard-Oxford atlas (Desikan et al., 2006). All voxel coordinates are reported in Montreal Neurological Institute (MNI) space.

To compute the main effect in a given ROI, we averaged the neural coefficients (betas) within a sphere of radius 10 mm centered at the peak voxel of the ROI for each subject, and then performed a two-tailed t-test against 0 with the averaged betas across subjects.

Following previous authors (Badre et al., 2012), we focus the rest of our analyses on positive clusters only. Negative clusters would indicate activity that is inversely correlated with RU and TU, possibly resulting from active neuronal inhibition (Wade, 2002). It is plausible that neurons in these clusters encode uncertainty by suppressing their activity, which could also be understood as positive encoding of certainty. However, this interpretation is inconsistent with

a rate code implementation of our hybrid model, particularly in the neocortex where most neurons are excitatory projection neurons and where all of our ROIs reside.

Decoding

If the brain regions identified by the GLM encode the subject's trial-by-trial estimates of RU_t and TU_t and those estimates dictate choices, as our model predicts, then we should be able to read out those subjective estimates and use them to improve our predictions for the subject's choices. This can be achieved by "inverting" the GLM and solving for RU_t and TU_t based on the neural data y_t, the beta coefficients β, and the other regressors X. Using ridge regression to prevent overfitting, the subjective estimate for RU for a given voxel on trial t can be computed as:

$$|\widehat{\mathrm{RU}}| = \left(y - \sum_{i:x_i \neq |\mathrm{RU}|} x_i \beta_i \right) \beta_{|\mathrm{RU}|} / (\beta_{|\mathrm{RU}|}^2 + \lambda) \tag{9}$$

where y is the neural signal, x_i is any regressor other than $|\mathrm{RU}|$, β_i is the corresponding beta coefficient computed by SPM, $\beta_{|\mathrm{RU}|}$ is the beta coefficient for $|\mathrm{RU}|$, and λ is the ridge regularization constant (voxel and trial indices were omitted to keep the notation uncluttered). To account for the hemodynamic lag, we approximated the neural activity at time t as the raw BOLD signal at time $t+5$ s, which corresponds to the peak of the canonical HRF in SPM (spm_hrf). Since the beta coefficients were already fit by SPM, we could not perform cross-validation to choose λ and so we arbitrarily set $\lambda = 1$. To remain consistent with SPM, we also whitened the neural data y and removed filter confounds (spm_filter).

To obtain a signed estimate for RU, we flipped the sign based on the model-based RU:

$$\widehat{\mathrm{RU}} = \begin{cases} |\widehat{\mathrm{RU}}| & \text{if RU} >= 0 \\ -|\widehat{\mathrm{RU}}| & \text{if RU} < 0 \end{cases} \tag{10}$$

Note that our goal is to test whether we can improve our choice predictions, given that we already know the model-based

RU. Finally, we averaged $\widehat{\mathrm{RU}}$ across all voxels within the given ROI (a 10 mm sphere around the peak voxel) to obtain a single trial-by-trial subjective estimate $\widehat{\mathrm{RU}}$ for that ROI. We used the same method to obtain a trial-by-trial subjective estimate of $\widehat{\mathrm{TU}}$.

To check if the neurally-derived $\widehat{\mathrm{RU}}$ predicts choices, we augmented the probit regression model of choice in Eq. 4 to:

$$P(a_t = \text{left}|\mathbf{w}) = \Phi(w_1 \mathrm{V}_t + w_2 \mathrm{RU}_t + w_3 \mathrm{V}_t/\mathrm{TU}_t + w_4 \widehat{\mathrm{RU}}_t), \tag{11}$$

where we first orthogonalized $\widehat{\mathrm{RU}}_t$ with respect to RU_t using Gram-Schmidt orthogonalization (spm_orth), ensuring that any contribution of $\widehat{\mathrm{RU}}_t$ to choice predictions is above and beyond what RU_t already contributes. After fitting the model using maximum-likelihood estimation (fitglme), we compared it to the original model (Eq. 4) using a likelihood ratio test (compare in MATLAB) and report the probability (p-value) that choices were generated by the original (null) model and not the augmented model.

Similarly, we defined an augemented model for $\widehat{\mathrm{TU}}$:

$$P(a_t = \text{left}|\mathbf{w}) = \Phi(w_1 \mathrm{V}_t + w_2 \mathrm{RU}_t + w_3 \mathrm{V}_t/\mathrm{TU}_t + w_5 \mathrm{V}_t/\widehat{\mathrm{TU}}_t), \tag{12}$$

where $\mathrm{V}_t/\widehat{\mathrm{TU}}_t$ was first orthogonalized with respect to $\mathrm{V}_t/\mathrm{TU}_t$. We similarly compared this model to the original model to check if it improves choice predictions.

Finally, once we identified a single ROI for each type of uncertainty using the above analyses and the cross-subject analysis (below), we entered both subjective estimates from the corresponding ROIs into the same augmented model:

$$P(a_t = \text{left}|\mathbf{w}) = \Phi(w_1 \mathrm{V}_t + w_2 \mathrm{RU}_t + w_3 \mathrm{V}_t/\mathrm{TU}_t + w_4 \widehat{\mathrm{RU}}_t + w_5 \mathrm{V}_t/\widehat{\mathrm{TU}}_t), \tag{13}$$

We then compared this model to the single-estimate augmented models (Eq. 11 and Eq. 12) using the same methods.

Cross-Subject Analysis

We hypothesized that cross-subject variability in exploration strategy would be reflected in the strength of neural encoding of the corresponding uncertainty regressor. In particular, subjects with weak or noisy representations of RU (smaller $\beta_{|\mathrm{RU}|}$) would make choices that are less consistent with directed exploration (smaller w_2 from the probit regression in Eq. 4). We thus performed a Pearson correlation across subjects between the neural coefficient $\beta_{|\mathrm{RU}|}$, averaged within a given ROI, and the random effects behavioral coefficient w_2. Similarly, we correlated β_{TU} with w_3.

References

Auer P. (2002). Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research *3*, 397–422.

Badre D., Doll B.B., Long N.M., and Frank M.J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. Neuron *73*, 595–607.

Beharelle A.R., Polanía R., Hare T.A., and Ruff C.C. (2015). Transcranial stimulation over frontopolar cortex elucidates the choice attributes and neural mechanisms used to resolve exploration–exploitation trade-offs. Journal of Neuroscience *35*, 14544–14556.

Blanchard T.C., and Gershman S.J. (2018). Pure correlates of exploration and exploitation in the human brain. Cognitive, Affective, & Behavioral Neuroscience *18*, 117–126.

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.

Busemeyer J.R., and Townsend J.T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychological review *100*, 432.

Chapelle O., and Li L. (2011). An empirical evaluation of Thompson sampling. In Advances in neural information processing systems. pp. 2249–2257.

Cohen J.D., McClure S.M., and Angela J.Y. (2007). Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. Philosophical Transactions of the Royal Society of London B: Biological Sciences *362*, 933–942.

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.

Desikan R.S., Ségonne F., Fischl B., Quinn B.T., Dickerson B.C., Blacker D., Buckner R.L., Dale A.M., Maguire

R.P., Hyman B.T., Albert M.S., and Killiany R.J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage *31*, 968–980.

Dezza I.C., Angela J.Y., Cleeremans A., and Alexander W. (2017). Learning the value of information and reward over time when solving exploration-exploitation problems. Scientific reports *7*, 16919.

Eickhoff S.B., Stephan K.E., Mohlberg H., Grefkes C., Fink G.R., Amunts K., and Zilles K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. NeuroImage *25*, 1325–1335.

Erev I., and Barron G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. Psychological review *112*, 912.

Feinberg D.A., Moeller S., Smith S.M., Auerbach E., Ramanna S., Gunther M., Glasser M.F., Miller K.L., Uĝurbil K., and Yacoub E. (2010). Multiplexed echo planar imaging for sub-second whole brain fMRI and fast diffusion imaging. PLoS ONE *5*, e15710.

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. Nature neuroscience *12*, 1062.

Gershman S.J. (2018a). Deconstructing the human algorithms for exploration. Cognition *173*, 34–42.

Gershman S.J. (2018b). Uncertainty and exploration. bioRxiv p. 265504.

Gershman S.J., and Tzovaras B.G. (2018). Dopaminergic genes are associated with both directed and random exploration. Neuropsychologia *120*, 97–104.

Hertwig R., Barron G., Weber E.U., and Erev I. (2004). Decisions from experience and the effect of rare events in risky choice. Psychological science *15*, 534–539.

May B.C., Korda N., Lee A., and Leslie D.S. (2012). Optimistic Bayesian sampling in contextual-bandit problems. Journal of Machine Learning Research *13*, 2069–2106.

Mehlhorn K., Newell B.R., Todd P.M., Lee M.D., Morgan K., Braithwaite V.A., Hausmann D., Fiedler K., and Gonzalez C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. Decision *2*, 191.

Menon V., and Uddin L.Q. (2010). Saliency, switching, attention and control: a network model of insula function. Brain Structure and Function *214*, 655–667.

Moeller S., Yacoub E., Olman C.A., Auerbach E., Strupp J., Harel N., and Uĝurbil K. (2010). Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. Magn Reson Med *63*, 1144–1153.

Mumford J., Poline J.B., and Poldrack R. (2015). Orthogonalization of regressors in fMRI models. PLoS ONE *10*, e0126255.

Myers J.L., and Sadler E. (1960). Effects of range of payoffs as a variable in risk taking. Journal of Experimental Psychology *60*, 306.

Ohira H., Matsunaga M., Murakami H., Osumi T., Fukuyama S., Shinoda J., and Yamada J. (2013). Neural mechanisms mediating association of sympathetic activity and exploration in decision-making. Neuroscience *246*, 362–374.

Peirce J.W. (2007). PsychoPy - Psychophysics software in Python. J Neurosci Methods *162*, 8–13.

Rolls E.T., Joliot M., and Tzourio-Mazoyer N. (2015). Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. NeuroImage *122*, 1 – 5.

Schulz E., and Gershman S.J. (in press). The algorithmic architecture of exploration in the human brain. Current Opinion in Neurobiology .

Schulz E., Konstantinidis E., and Speekenbrink M. (2015). Learning and decisions in contextual multi-armed bandit tasks. In CogSci.

Somerville L.H., Sasse S.F., Garrad M.C., Drysdale A.T., Abi Akar N., Insel C., and Wilson R.C. (2017). Charting the expansion of strategic exploratory behavior during adolescence. Journal of Experimental Psychology: General *146*, 155–164.

Speekenbrink M., and Konstantinidis E. (2015). Uncertainty and exploration in a restless bandit problem. Topics in cognitive science *7*, 351–367.

Srinivas N., Krause A., Kakade S., and Seeger M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In Proceedings of the 27th International Conference on International Conference on Machine Learning (USA: Omnipress), ICML'10, pp. 1015–1022.

Thompson W.R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika *25*, 285–294.

Tomov M.S., Dorfman H.M., and Gershman S.J. (2018). Neural computations underlying causal structure learning. Journal of Neuroscience *38*, 7143–7157.

Tzourio-Mazoyer N., Landeau B., Papathanassiou D., Crivello F., Etard O., Delcroix N., Mazoyer B., and Joliot M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage *15*, 273–289.

van der Kouwe A.J., Benner T., Salat D.H., and Fischl B. (2008). Brain morphometry with multiecho MPRAGE. NeuroImage *40*, 559–569.

Wade A.R. (2002). The negative bold signal unmasked. Neuron *36*, 993–995.

Weber E.U., Shafir S., and Blais A.R. (2004). Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. Psychological review *111*, 430.

Wilson R.C., Geana A., White J.M., Ludvig E.A., and Cohen J.D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. Journal of Experimental Psychology: General *143*, 2074.

Xu J., Moeller S., Auerbach E.J., Strupp J., Smith S.M., Feinberg D.A., Yacoub E., and Uĝurbil K. (2013). Evaluation of slice accelerations using multiband echo planar imaging at 3 T. NeuroImage *83*, 991–1001.

Yechiam E., and Busemeyer J.R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. Psychonomic bulletin & review *12*, 387–402.

Zajkowski W.K., Kossut M., and Wilson R.C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. eLife *6*, e27430.

Supplemental Figures and Tables

regressor	**event**	**duration**	**pmods**	**which trials**
trial_onset	trial onset	0 s	$\lvert RU_t \rvert$, TU_t, t	non-timeout
trial_onset_timeout	trial onset	0 s		timeout
trial_onset_chose_1	trial onset	0 s		chose arm 1
button_press	reaction time	0 s		all
feedback_onset	feedback onset	0 s		all

Table S1. **GLM** used to analyze the fMRI data.

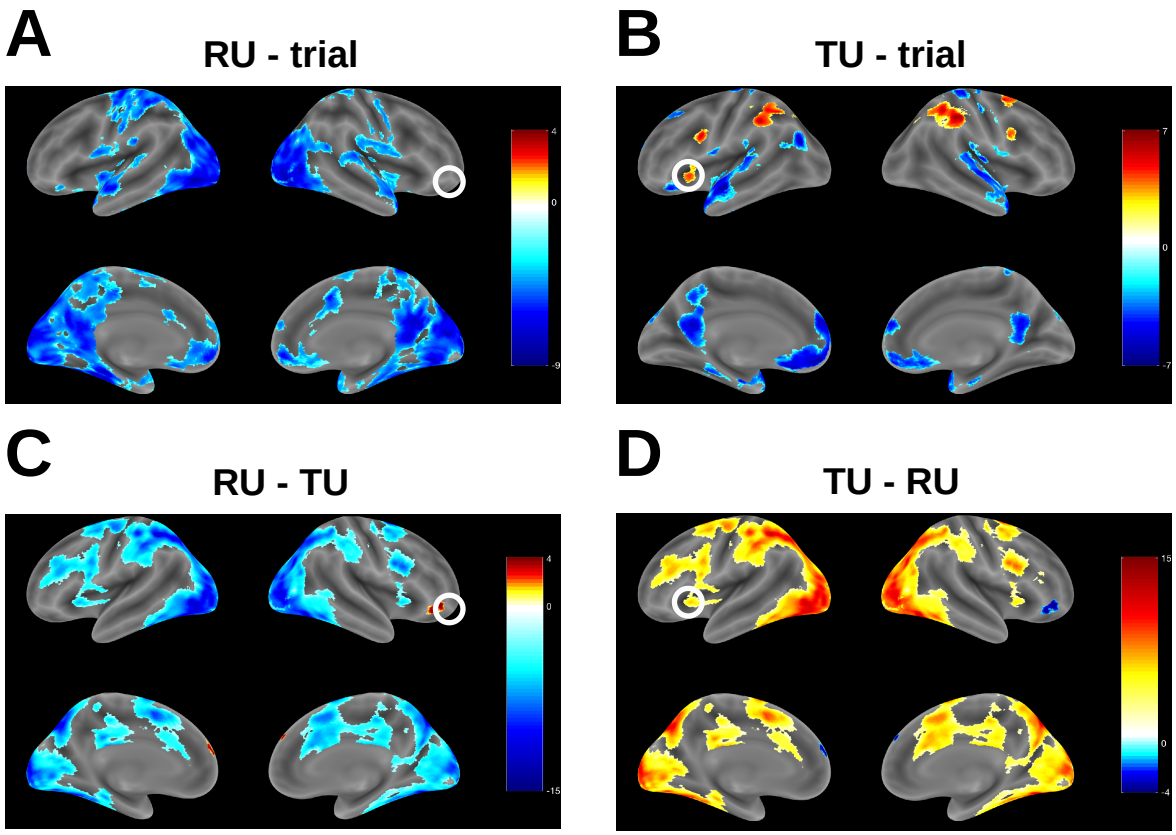

Figure S1. **Corrected whole-brain contrasts** with single voxels thresholded at $p < 0.001$ and cluster FWE correction applied at significance level $\alpha = 0.05$.

(A) RU - trial. The circled ROI in right RLPFC (MNI 34 50 -10) was used in the analyses in Figure 3.

(B) TU - trial. The circled ROI in left anterior insula RLPFC (MNI -30 22 2) was used in the analyses in Figure 4.

(C) RU - TU. The circled ROI in right RLPFC (MNI 34 50 -10) was used in the analyses in Figure 3.

(D) TU - RU. The circled ROI in left anterior insula RLPFC (MNI -30 22 2) was used in the analyses in Figure 4.

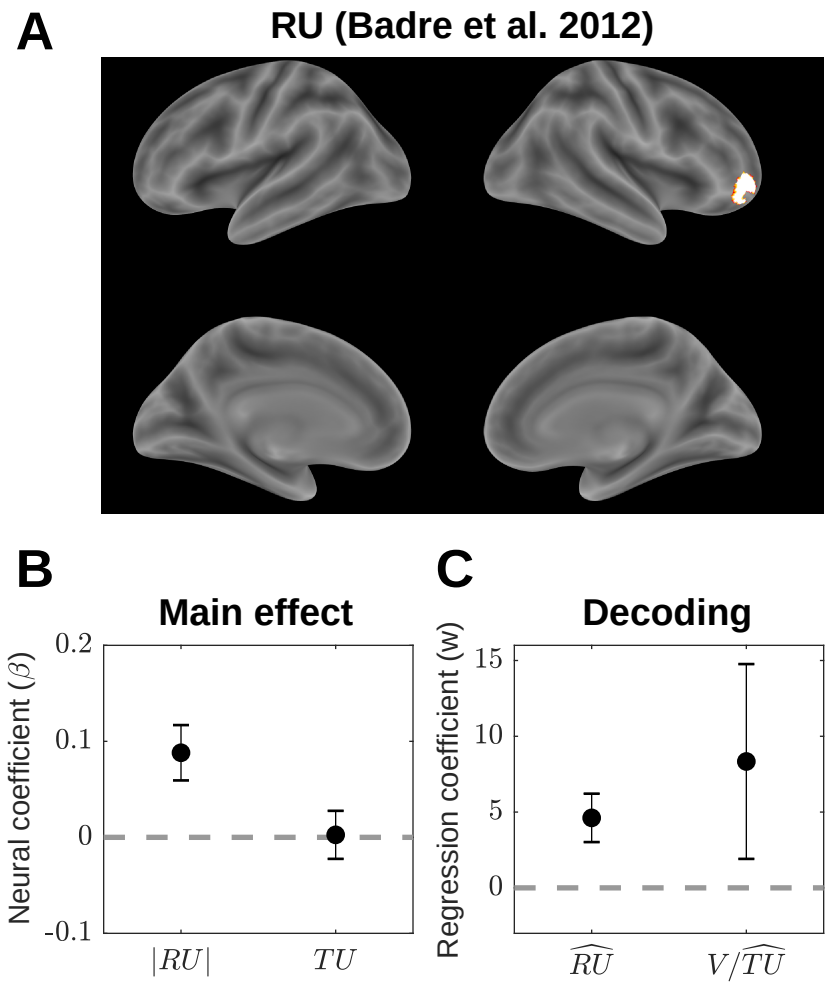

Figure S2. **Right RLPFC encodes RU but not TU.**

(A) Right RLPFC ROI (MNI 36 56 -8) associated with relative uncertainty by Badre et al. (2012) (10 mm sphere around peak voxel).

(B) and (C) same as in Figure 3 but using the right RLPFC ROI from Badre et al. (2012).

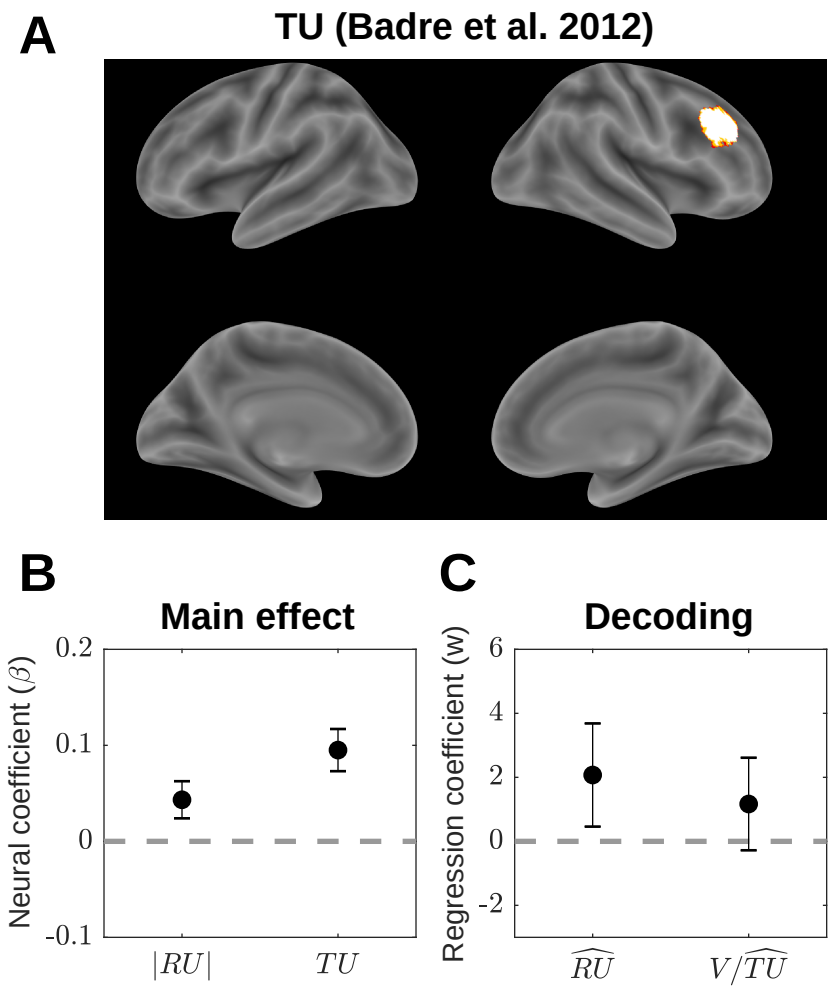

Figure S3. **No specific encoding of TU in DLPFC.**

(A) Right DLPFC ROI (MNI 40 30 34) associated with mean uncertainty by Badre et al. (2012) (10 mm sphere around voxel).

(B) and (C) same as in Figure 3 but using the right DLPFC ROI from Badre et al. (2012).

Sign	**Brain region**	**BA**	**Extent**	***t*-value**	**MNI coord.**
Positive	Middle frontal gyrus (R)	47	102	4.922	22 36 -2
	Middle frontal gyrus (R)	47	177	4.721	34 50 -10
Negative	*Inferior temporal gyrus (L)	37	71030	-3.808	-46 -58 -10
	*Inferior temporal gyrus (L)	20	71030	-9.129	-40 -30 -22
	Superior frontal gyrus, dorsolateral (L)	9	235	-5.097	-28 38 40

Table S2. **RU - trial (uncorrected)**. Brain regions in which the BOLD signal tracks RU above and beyond trial number (corresponding to Figure 3A). Anatomical labels and MNI coordinates are based on the peak voxel (maximum t-statistic) from each cluster. Single voxels were thresholded at $p < 0.001$ and clusters were thresholded at extent >= 100 voxels. Regions were labeled using the Automated Anatomical Labeling (AAL2) atlas. MNI = Montreal Neurological Institute; BA = Brodmann area. * - survives FWE correction at significance level $\alpha = 0.05$.

Sign	Brain region	BA	Extent	*t*-value	MNI coord.
Positive	*Inferior parietal gyrus (R)	40	2042	7.508	38 -46 44
	*Inferior parietal gyrus (L)	40	1521	7.298	-46 -40 48
	*Superior frontal gyrus (R)		468	6.231	22 -2 52
	Superior frontal gyrus (L)		179	5.793	-26 -4 44
	*Precentral gyrus (L)	44	282	5.791	-42 2 28
	*Precentral gyrus (R)	44	258	4.841	46 8 30
	*Insula (L)	47	268	4.815	-30 22 2
	Middle frontal gyrus (R)	46	183	4.728	38 30 34
	Middle frontal gyrus (L)	45	137	4.149	-44 30 32
Negative	*Middle temporal gyrus (L)	20	4376	-7.801	-54 -10 -14
	*Gyrus rectus (L)	11	4161	-7.561	-4 22 -22
	*Angular gyrus (L)	39	1000	-6.958	-46 -74 28
	*Parahippocampal gyrus (R)		4429	-6.905	16 -4 -30
	*Precuneus (L)	30	2344	-6.773	-6 -54 14
	*Middle frontal gyrus (R)	6	310	-5.649	54 -6 52
	*Postcentral gyrus (R)	4	976	-5.644	28 -30 66
	Cerebellum (R)		208	-5.421	12 -50 -38
	Posterior orbital gyrus (R)	47	101	-5.392	36 30 -20
	*Superior frontal gyrus (L)		303	-5.205	-22 18 36
	*Paracentral lobule (L)	6	302	-5.178	-18 -26 72
	*Fusiform gyrus (L)	37	305	-5.010	-32 -36 -16
	Lingual gyrus (L)	17	130	-4.892	-4 -80 -10
	IFG pars triangularis (L)	45	146	-4.550	-54 30 0

Table S3. **TU - trial (uncorrected)**. Brain regions in which the BOLD signal tracks TU above and beyond trial number (corresponding to Figure 4A). Notation and procedures as in Table S2.

model	DF	AIC	BIC	LL	LR-stat	p-value
Hybrid model	3	7502.54	7524.01	-3748.27		
Augmented $\widehat{\mathrm{RU}}$ model	4	7499.50	7528.13	-3745.75	5.04	p = 0.025

Table S4. **$\widehat{\mathrm{RU}}$ decoded from right RLPFC predicts choices.** Results from likelihood ratio test comparing the hybrid model (Eq. 4) with the augmented model which includes the neurally decoded $\widehat{\mathrm{RU}}$ (Eq. 11) from right RLPFC (MNI 34 50 -10). DF = degrees of freedom; AIC = Akaike information criterion; BIC = Bayesian information criterion; LL = maximized log likelihood; LR-stat = likelihood ratio test statistic.

Brain region	MNI coord.	AIC	BIC	LL	LR-stat	p-value (uncorr.)	p-value (corr.)
Inferior parietal gyrus (R)	38 -46 44	7502.07	7530.70	-3747.03	2.47	$p = 0.1160$	$p = 0.6704$
Inferior parietal gyrus (L)	-46 -40 48	7504.53	7533.15	-3748.26	0.01	$p = 0.9102$	$p = 1.0000$
Superior frontal gyrus (R)	22 -2 52	7496.13	7524.76	-3744.07	8.41	$*p = 0.0037$	$*p = 0.0332$
Superior frontal gyrus (L)	-26 -4 44	7503.34	7531.97	-3747.67	1.20	$p = 0.2742$	$p = 0.9441$
Precentral gyrus (L)	-42 2 28	7504.08	7532.70	-3748.04	0.46	$p = 0.4958$	$p = 0.9979$
Precentral gyrus (R)	46 8 30	7485.03	7513.66	-3738.51	19.51	$*p < 10^{-4}$	$*p < 10^{-4}$
Insula (L)	-30 22 2	7486.67	7515.30	-3739.34	17.87	$*p < 10^{-4}$	$*p = 0.0002$
Middle frontal gyrus (R)	38 30 34	7503.46	7532.09	-3747.73	1.08	$p = 0.2988$	$p = 0.9590$
Middle frontal gyrus (L)	-44 30 32	7498.25	7526.87	-3745.12	6.29	$p = 0.0121$	$p = 0.1039$

Table S5. **$\widehat{\mathrm{TU}}$ decoded from left primary motor cortex, right premotor cortex, and left anterior insula predicts choices.** Results from likelihood ratio test comparing the hybrid model (Eq. 4) with the augmented model which includes the neurally-decoded $\widehat{\mathrm{TU}}$ (Eq. 12). Notation as in Table S4, with statistics included for the augmented $\widehat{\mathrm{TU}}$ model only. * - significant after Bonferroni correction with adjusted $\alpha = 0.05/9 = 0.0056$.

Brain region	**MNI coord.**	**Pearson's r**	**p-value (uncorr.)**	**p-value (corr.)**
Superior frontal gyrus (R)	22 -2 52	0.34	p = 0.065	p = 0.183
Precentral gyrus (R)	46 8 30	0.30	p = 0.096	p = 0.262
Insula (L)	-30 22 2	0.43	*p = 0.016	*p = 0.048

Table S6. **Differences in TU encoding predict differences in random exploration.** Results from between-subject Pearson correlation between the neural coefficient for TU (β_{TU}) and the extent to which they employ random exploration (w_3 in Eq. 4). * - significant after Bonferroni correction with adjusted $\alpha = 0.05/3 = 0.017$.