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.

Acknowledgments: This research was supported by the Office of Naval Research Science of Autonomy program (N00014-17-1-2984), the National Institutes of Health (award number 1R01MH109177), and the Toyota Corporation. This work involved the use of instrumentation supported by the NIH Shared Instrumentation Grant Program award 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 correlated with BOLD activity 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 correlated with activity in left dorsolateral prefrontal cortex 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. 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.

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, 2018). These strategies are more adaptive in nonstationary environments and come in two distinct flavors: *directed* and *random* exploration strategies.

Directed exploration strategies *direct* the agent's choices towards uncertain options, which is equivalent to 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).

Previous work (Gershman, 2018; 2019; Gershman and Tzovaras, 2018) has shown that directed exploration in humans is well captured by the upper confidence bound (UCB) algorithm (Auer, 2002), in which the uncertainty 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. While earlier studies favored value-based random exploration strategies such as softmax exploration, later work (Wilson et al., 2014; Gershman, 2018) has shown that random exploration in people is additionally 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, 2018). A study by Gershman (2019) 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 reflected in right RLPFC and total uncertainty is reflected in left dorsolateral prefrontal cortex (DLPFC). The neural signal in right RLPFC predicted cross-trial variability in directed but not random exploration, whereas the neural signal in left DLPFC predicted cross-trial variability in random but not directed 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 DLPFC 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, 2019). 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 remained fixed within blocks and 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 across different block conditions, 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 expected 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).

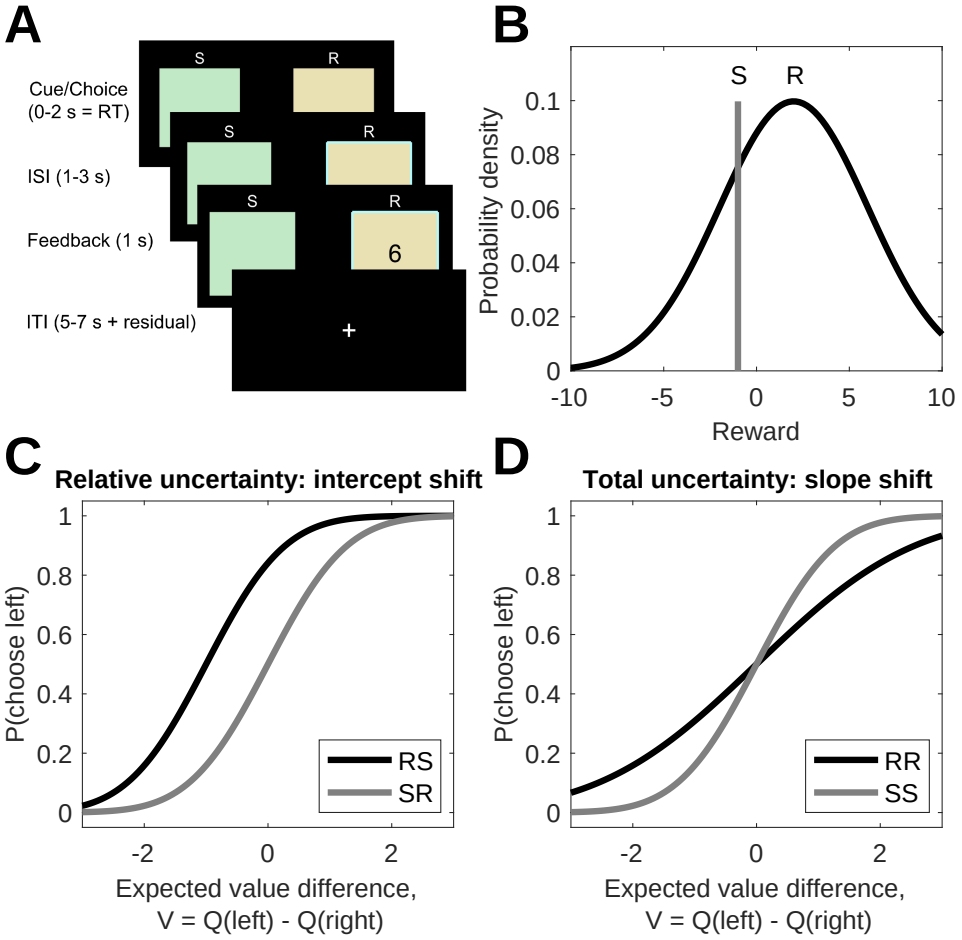

Figure 1. **Experimental Design and Predictions.**

(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.

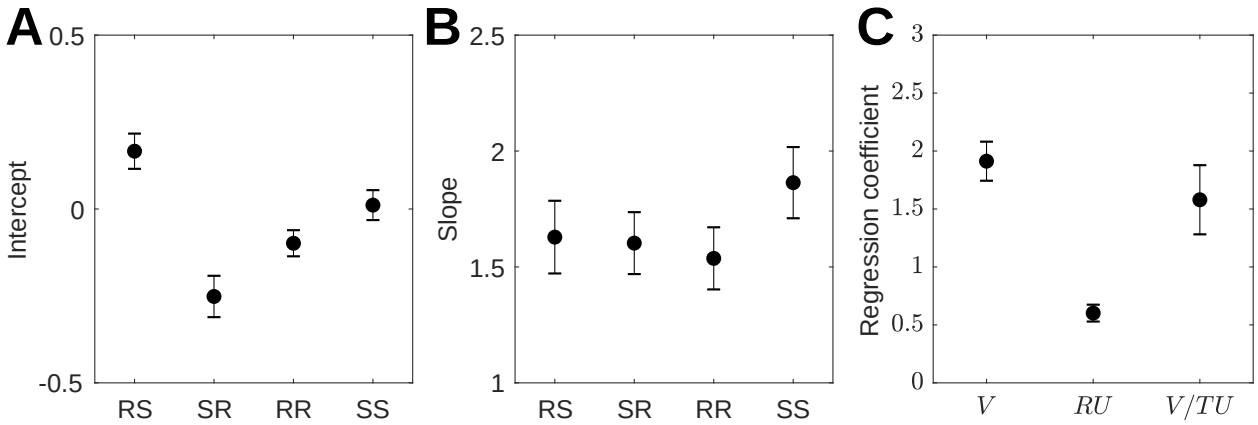

Figure 2. **Probit regression results.**

(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).

Overall, subjects identified the better option in each block (i.e., the option k with the greater expected reward $\mu(k)$) relatively quickly, with average performance plateauing by the middle of each block (Figure S2). Importantly, in accordance with the theory, we found that manipulating relative uncertainty (RS vs. SR) shifted the intercept of the choice probability function (Figure 2A, Figure S3): 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, Figure S3): 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, 2019; 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, 2019):

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

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

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

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 (2018) 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, 2019) 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 represented 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 the quantities used in computing the decision (Eq 4): absolute relative uncertainty ($|\text{RU}_t|$), total uncertainty (TU_t), absolute value difference ($|\text{V}_t|$), and absolute value difference scaled by total uncertainty ($|\text{V}_t|/\text{TU}_t$) as non-orthogonalized impulse regressors at trial onset (see Experimental Procedures). 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

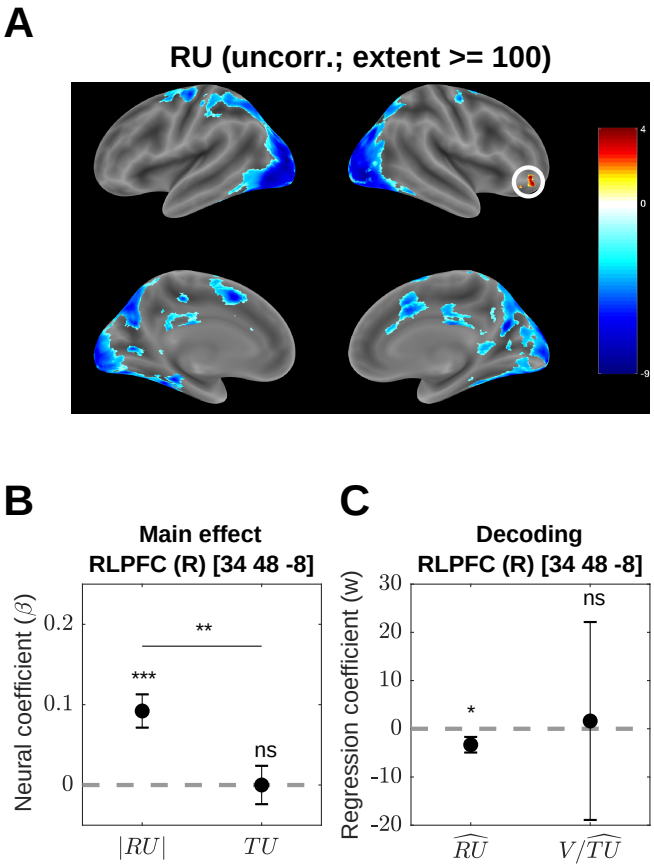

Figure 3. **Right RLPFC tracks RU but not TU.**

(A) Whole-brain RU contrast. 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 48 -8) was used in the subsequent plots (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. * - p < 0.05, ** - p < 0.01, *** - p < 0.001, **** - p < 0.0001.

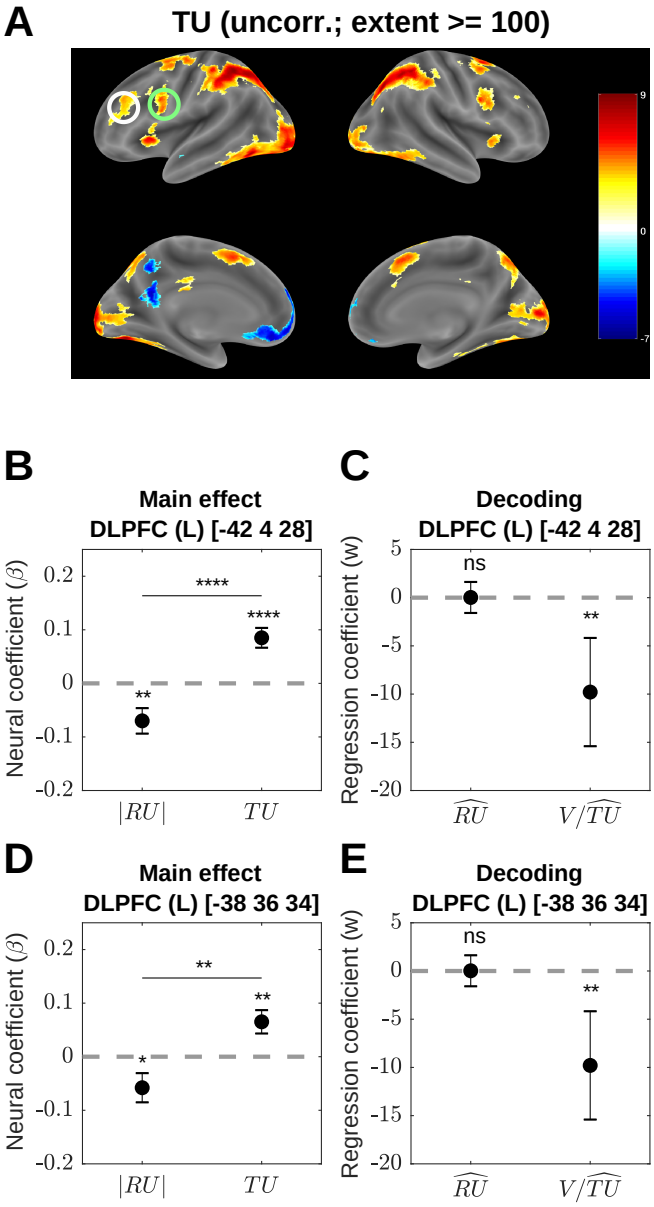

Figure 4. **Left DLPFC tracks TU but not RU.**

(A) Whole-brain TU contrast. 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 ROIs in left DLPFC (MNI -42 4 28, green; MNI -38 36 34, white) were used in subsequent plots (10 mm sphere around the peak voxel).

(B), (C) and (D), (E) same as in Figure 3 but using the left DLPFC ROIs.

threshold in order to identify candidate regions for our subsequent confirmatory analyses (corrected whole-brain results are reported in Figure S1).

For the RU contrast, we did not find any positive regions that survived FWE correction (Figure S1A). We then considered the uncorrected RU results with minimum cluster extent >= 100 voxels (Figure 3A and Table S2) and found a cluster in right RLPFC (MNI 34 48 -8) which showed a significant effect for RU ($t(30) = 4.5, p = 0.0001$) but not TU ($t(30) = 0.001, p = 0.99$; Figure 3B). The significance of this difference was confirmed by the RU - TU contrast within that ROI ($t(30) = 2.9, p = 0.007$, paired t-test). To remain consistent, we also considered the uncorrected TU contrast with minimum cluster extent >= 100 voxels (Figure 4 and Table S3; corrected results are shown in Figure S1B). We found bilateral positive clusters in the inferior parietal lobe, DLPFC, anterior insula, inferior temporal lobe, midcingulate cortex, superior posterior thalamus, and premotor cortex. All of these ROIs showed a negative effect for RU ($t(30) < 0$ for all ROIs), which was significant for inferior parietal gyrus (MNI -42 -40 52, $p = 0.001$, Bonferroni corrected for 10 comparions), thalamus (MNI -4 -16 14, $p = 0.03$), midbrain (MNI -8 -20 -12, $p = 0.03$), and midcingulate cortex (MNI -2 -26 30, $p = 0.0005$). We did not find any clusters for V or V/TU. We therefore used the regions of interest (ROIs) from the RU and TU contrasts in our subsequent analysis.

Subjective estimates of relative and total uncertainty predict choices

If the ROIs identified by the RU and TU contrasts encode relative and total uncertainty, respectively, 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 (MNI 34 48 -8; Table S4 and Figure 3C; $p = 0.04$, likelihood ratio test). Importantly, decoding $\widehat{\text{TU}}$ from right RLPFC and augmenting the model with $V/\widehat{\text{TU}}$ (Eq. 12) did not improve choice predictions ($p > 0.9$).

Similarly, we found a significant effect when augmenting the hybrid model with $V/\widehat{\text{TU}}$ (Eq. 12) when $\widehat{\text{TU}}$ was decoded from one of the two ROIs in DLPFC (MNI -42 4 28, $p = 0.006$; MNI -38 36 34, $p = 0.006$, likelihood ratio test, Bonferroni corrected for 10 comparisons; Table S5 and Figure 4C). Importantly, decoding $\widehat{\text{RU}}$ from the ROIs from the TU contrast did not improve choice predictions ($p > 0.2$ for all 10 ROIs). The two DLPFC ROIs were further validated by the TU - RU contrast, which was significant in both ROIs (respectively, $t(30) = 5.2, p = 0.00001$ and $t(30) = 3.3, p = 0.002$, paired t-tests). Together, these results show that variability in the neural representations of uncertainty in the corresponding regions predicts choices, consistent with the idea that those representations are used in a downstream decision computation.

We additionally confirmed that including $\widehat{\text{RU}}$ decoded from right RLPFC (MNI 34 48 -8) together with $V/\widehat{\text{TU}}$ decoded from either of the two DLPFC ROIs (MNI -42 4 28 and -38 36 34) in the same model (Eq. 13) improved choice predictions beyond the improvement from including either estimate alone (MNI -42 4 28: $LRStat = 12.1, p = 0.0005$ when comparing with $\widehat{\text{RU}}$ alone, Eq. 11, and $LRStat = 5.0, p = 0.03$ when comparing with $V/\widehat{\text{TU}}$ alone, Eq. 12; MNI -38 36 34: $LRStat = 11.9, p = 0.0006$ when comparing with $\widehat{\text{RU}}$ alone, Eq. 11, and $LRStat = 4.2, p = 0.04$ when comparing with $V/\widehat{\text{TU}}$ alone, Eq. 12, likelihood ratio test). Finally, to validate our choice of ROIs based on the whole-brain analysis, we repeated all of the analyses with the ROIs reported by Badre et al. (2012), which yielded the same qualitative results for relative uncertainty but not for total uncertainty (see Supplemental Information for details).

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. Similarly, our results point to left DLPFC as the neural locus of the subjective estimate of total uncertainty and random exploration.

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 activity in right RLPFC tracks the 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 definition of uncertainty as the posterior standard deviation of the mean (sometimes referred to as estimation uncertainty or ambiguity) is different from the expected uncertainty due to the unpredictability of the reward from the risky option on any particular trial (sometimes referred to as irreducible uncertainty or risk; Payzan-LeNestour and Bossaerts, 2011). These two forms of uncertainty are generally related, and in particular in our study, estimation uncertainty is higher, on average, for risky arms due to the variability of their rewards, which makes it difficult to estimate the mean exactly. However, they are traditionally associated with opposite behaviors: risk-aversion predicts that people would prefer the safe over the risky option (Kahneman and Tversky, 1979), while uncertainty-guided exploration predicts a preference for the risky option, all else equal (Figure 1C). While we did not seek to explicitly disentangle risk from estimation uncertainty in our study, our behavioral (Figure 2A and Figure S3) and neural (Figure 3) results are consistent with the latter interpretation.

Previous studies that have examined the neural correlates of uncertainty defined as risk have identified some of the regions we found for total uncertainty (Figure 4 and Table S3), such as anterior insula (Preuschoff et al., 2008), DLPFC, parietal cortex (Mohr et al., 2010), and dACC (Christopoulos et al., 2009). In particular, the posterior region we identified in left DLPFC (green circle in Figure 4A) has been associated with perceptual uncertainty (Daniel et al., 2011; Nastase et al., 2014), uncertainty due to changes in environmental stability (Kluger and Schubotz, 2017), stimulus entropy (Fan et al., 2014), and decision under risk (Brevers et al., 2015; Causse et al., 2013). The anterior region we identified in right DLPFC (white circle in Figure 4A) has also been shown to track risk (Tobler et al., 2007). Our findings suggest that left DLPFC might be additionally tracking total estimation uncertainty, which underlies the random exploration component of choices in our model. Future work could disentangle differential contribution of risk versus estimation uncertainty to neural activity and choices.

Our results are broadly consistent with results reported by Badre et al. (2012) who identified neural correlates of relative uncertainty in RLPFC and neural correlates of total uncertainty in DLPFC. However, our results differ in the lateralization of total uncertainty: Badre et al. (2012)'s work suggests that total uncertainty is encoded in right DLPFC, while our results point to left DLPLC. We also did not find an effect of total uncertainty in the right DLPFC ROI they reported (see Supplemental Information). Further, our decoding analysis goes beyond Badre et al. (2012)'s work, providing strong support for the hypothesis that variability in the neural signals in these regions reflects variability in the underlying uncertainty computations.

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 DLPFC 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 correlated with activity in right RLPFC and influenced directed exploration, while total uncertainty was correlated with activity in left DLPFC and influenced random exploration. Subjective trial-by-trial estimates decoded from both regions predicted subject responding, suggesting 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 (2019). 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 (2019), 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, 2018). In order to prevent degeneracy of the Kalman update for safe arms, we used $\tau^2(S) = 0.00001$ instead of zero, which is equivalent to assuming a negligible amount of noise even for safe arms. Notice that in this case, the learning rate is $\alpha_t \approx 1$ and as soon as the safe arm k is sampled, the posterior mean and variance are updated to $Q_{t+1}(k) \approx \mu(k)$ and $\sigma_t^2(k) \approx 0$, respectively.

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 (i.e., all trials on which the subject made a choice within 2 s of trial onset). We confirmed the parameter recovery capabilities of our approach by running the hybrid model genaratively on the same bandits as the subjects (Wilson and Collins, 2019). We drew the weights in each simulation according to $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, 10 \times \mathbf{I})$ and repeated the process 1000 times. We found a strong correlation between the generated and the recovered weights (Figure S4A; $r > 0.99, p < 10^{-8}$ for all weights) and no correlation between the recovered weights (Figure S4B; $r < 0.03, p > 0.3$ for all pairs), thus validating our approach.

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, $|\mathrm{RU}_t|$, total uncertainty, TU_t, absolute value difference, $|\mathrm{V}_t|$, and

absolute value difference scaled by total uncertainty, $|V_t|/TU_t$, as parametric modulators for an impulse regressor at trial onset (trial_onset). Following (Badre et al., 2012), we used $|RU_t|$ instead of RU_t to account for our arbitrary choice of arm 1 and arm 2 (note that TU_t is always positive). We used the absolute value difference $|V_t|$ for the same reason. The trial_onset regressor was only included on trials on which the subject responded within 2 s of trial onset (i.e., non-timeout trials). We included a separate regressor at trial onset for trials on which the subject timed out (i.e., failed to respond within 2 s of trial onset) 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.

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. To compute a contrast in a given ROI, we performed a paired two-tailed t-test across subjects between the betas for one regressor (e.g. RU) and the betas for the other (e.g. TU), averaged within a 10

mm-radius sphere.

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.

Note that this analysis cannot be circular since it evaluates the ROIs based on behavior, which was not used to define the neural GLM (Kriegeskorte et al., 2009). In particular, all regressors and parametric modulators in the GLM were

defined purely based on model-derived quantities; we only take into account subjects' choices when fitting the weights $\mathbf{w}$ (Eq. 4), which were not used to define the GLM. Furthermore, since all model-derived regressors are also included in all augmented models of choice (Eq. 11, 12, 13), any additional choice information contributed by the neurally decoded regressors is necessarily above and beyond what was already included in the neural GLM.

Finally, once we identified a single ROI for each type of uncertainty using the above analyses, we entered both subjective estimates from the corresponding ROIs into the same augmented model:

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

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

References

Alexander W.H., and Brown J.W. (2011). Medial prefrontal cortex as an action-outcome predictor. Nature neuroscience *14*, 1338.

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.

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.

Brevers D., Bechara A., Hermoye L., Divano L., Kornreich C., Verbanck P., and Noël X. (2015). Comfort for uncertainty in pathological gamblers: A fmri study. Behavioural brain research *278*, 262–270.

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.

Causse M., Péran P., Dehais F., Caravasso C.F., Zeffiro T., Sabatini U., and Pastor J. (2013). Affective decision making under uncertainty during a plausible aviation task: An fmri study. NeuroImage *71*, 19–29.

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

Christopoulos G.I., Tobler P.N., Bossaerts P., Dolan R.J., and Schultz W. (2009). Neural correlates of value, risk, and risk aversion contributing to decision making under risk. Journal of Neuroscience *29*, 12574–12583.

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.

Daniel R., Wagner G., Koch K., Reichenbach J.R., Sauer H., and Schlösser R.G. (2011). Assessing the neural basis of uncertainty in perceptual category learning through varying levels of distortion. Journal of Cognitive Neuroscience *23*, 1781–1793.

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.

Fan J., Van Dam N.T., Gu X., Liu X., Wang H., Tang C.Y., and Hof P.R. (2014). Quantitative characterization of functional anatomical contributions to cognitive control under uncertainty. Journal of cognitive neuroscience *26*, 1490–1506.

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

Gershman S.J. (2019). Uncertainty and exploration. Decision p. 265504.

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

Grabenhorst F., and Rolls E.T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. Trends in cognitive sciences *15*, 56–67.

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.

Kahneman D., and Tversky A. (1979). Prospect theory: An analysis of decision under risk. Econometrica *47*, 263–292.

Kluger D.S., and Schubotz R.I. (2017). Strategic adaptation to non-reward prediction error qualities and irreducible uncertainty in fmri. cortex *97*, 32–48.

Kriegeskorte N., Simmons W.K., Bellgowan P.S., and Baker C.I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. Nature neuroscience *12*, 535.

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.

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.

Mohr P.N., Biele G., and Heekeren H.R. (2010). Neural processing of risk. Journal of Neuroscience *30*, 6613–6619.

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.

Nastase S., Iacovella V., and Hasson U. (2014). Uncertainty in visual and auditory series is coded by modality-general and modality-specific neural systems. Human brain mapping *35*, 1111–1128.

Payzan-LeNestour E., and Bossaerts P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. PLOS Computational Biology *7*, 1–14.

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

Preuschoff K., Quartz S.R., and Bossaerts P. (2008). Human insula activation reflects risk prediction errors as well as risk. Journal of Neuroscience *28*, 2745–2752.

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.

Tobler P.N., O'Doherty J.P., Dolan R.J., and Schultz W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. Journal of neurophysiology *97*, 1621–1632.

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., and Collins A. (2019). Ten simple rules for the computational modeling of behavioral data .

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 Uğurbil 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 Information

A priori ROIs

A previous study by Badre et al. (2012) used a similar analysis to identify regions associated with relative and total uncertainty. Although their task and model were different than ours, the underlying uncertainty computations are mathematically equivalent. We therefore repeated our analyses with the ROIs from their study. We chose the ROIs that were identified using regressors analogous to our RU (right RLPFC, MNI 36 56 -8, Figure S6A) and TU (right DLPFC, MNI 40 30 34, Figure S7A).

We found that right RLPFC from their study was significantly correlated with RU ($t(30) = 3.3, p = 0.002$; Figure S6B) but not TU ($t(30) = -0.6, p > 0.5$) and that this difference was significant ($t(30) = 2.5, p = 0.02$). Furthermore, decoding $\widehat{\mathrm{RU}}$ from that ROI improved choice predictions (Figure S6C; $p = 0.02$, likelihood ratio test) but decoding $\widehat{\mathrm{TU}}$ did not ($p > 0.8$). These results are consistent with our right RLPFC ROI findings (MNI 34 48 -8, Figure 3A).

However, we did not find a significant effect of TU in right DLPFC ($t(30) = 0.9, p > 0.3$). Correspondingly, decoding $\widehat{\mathrm{TU}}$ from right DLPFC did not improve choice predictions ($p > 0.05$). These results are also consistent with our findings, which point to left DLPFC (MNI -42 4 28 and -38 36 34) rather than right DLPFC as the locus of TU.

Variance inflation factors

Since the parametric modulators of the trial_onset regressor are correlated (e.g. V and V/TU), we computed variance inflation factors (VIFs) for all four parametric modulators on all runs for all subjects (Figure S5). Overall, only 5% of VIFs for RU were above the threshold of 10 (12 out of 240), however we found that around 20% of VIFs for TU and V/TU were above 10 (53 and 47 out of 240, respectively), and more than half of VIFs for V were above 10 (128 out of 240). Note that inferences are valid even for regressors with a high VIF since the estimate of the beta coefficients is still unbiased and thus the type I error rate is preserved (Mumford et al., 2015). However, the inflated variance of the beta estimates might reduce the power of the analysis. To investigate this possibility, we re-analyzed the data using four new GLMs that were nearly identical to the original GLM, with the only difference that each GLM had a single

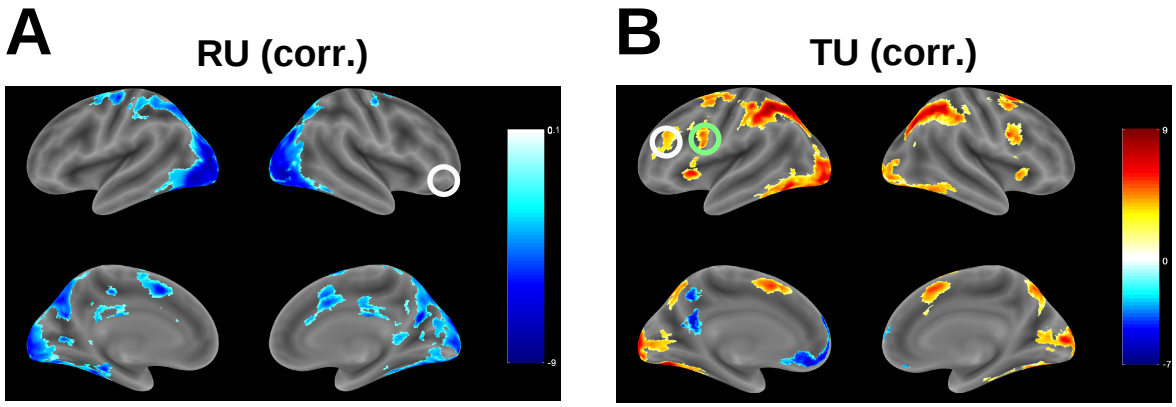

Figure S1. **Corrected whole-brain contrasts** with single voxels thresholded at $p < 0.001$ and cluster FWE correction applied at significance level $\alpha = 0.05$.

(A) RU contrast. The circled ROI in right RLPFC (MNI 34 48 -8) was used in the analyses in Figure 3.

(B) TU contrast. The circled ROIs in left DLPFC (MNI -42 4 28, green; MNI -38 36 34, white) were used in the analyses in Figure 4.

parametric modulator at trial onset rather than four. Thus there was one GLM with RU, one with TU, one with V, and one with V/TU.

We found the same network of brain regions for RU and TU (Figure S8A and B, respectively) and no regions for V/TU. For V, we found clusters in medial PFC (Figure S8C), which is consistent with previous reports of value coding in this region (Alexander and Brown, 2011; Grabenhorst and Rolls, 2011). Since the value-coding function of this region has been charactarized extensively (Daw et al., 2006) and since our primary interest was in the role of uncertainty in guiding exploration, we chose not to pursue this finding.

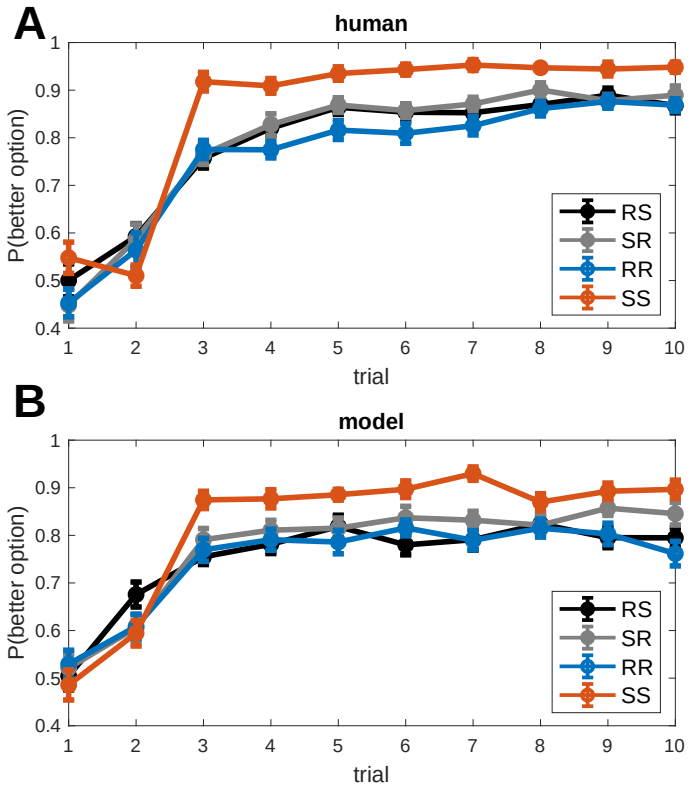

Figure S2. **Learning curves** for human (A) and model (B) data. The better option is defined as the option with the greater expected reward $\mu(k)$.

regressor	**event**	**duration**	**pmods**	**which trials**
trial_onset	trial onset	0 s	$\lvert RU_t \rvert$, TU_t, $\lvert V_t \rvert$, $\lvert V_t \rvert / TU_t$	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.

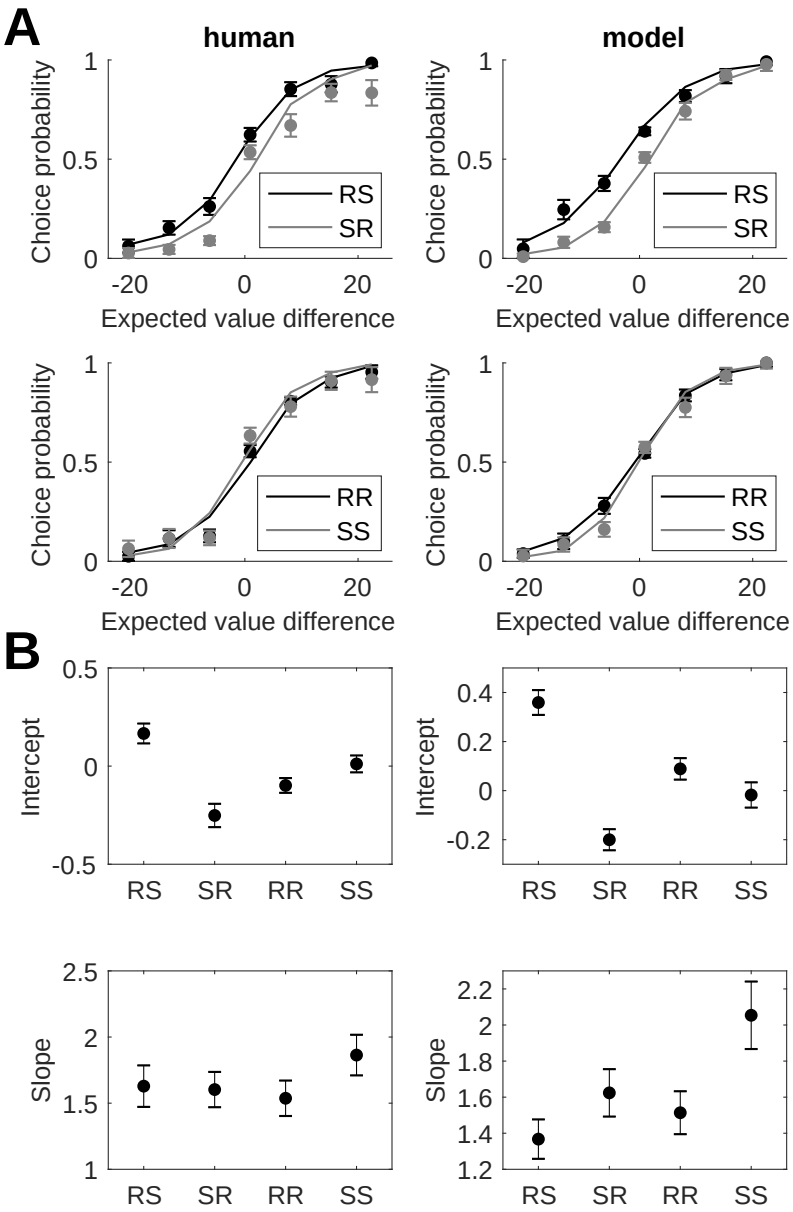

Figure S3. **Choice probability functions (A) and probit regression results (B)** for human (left) and model (right) data.

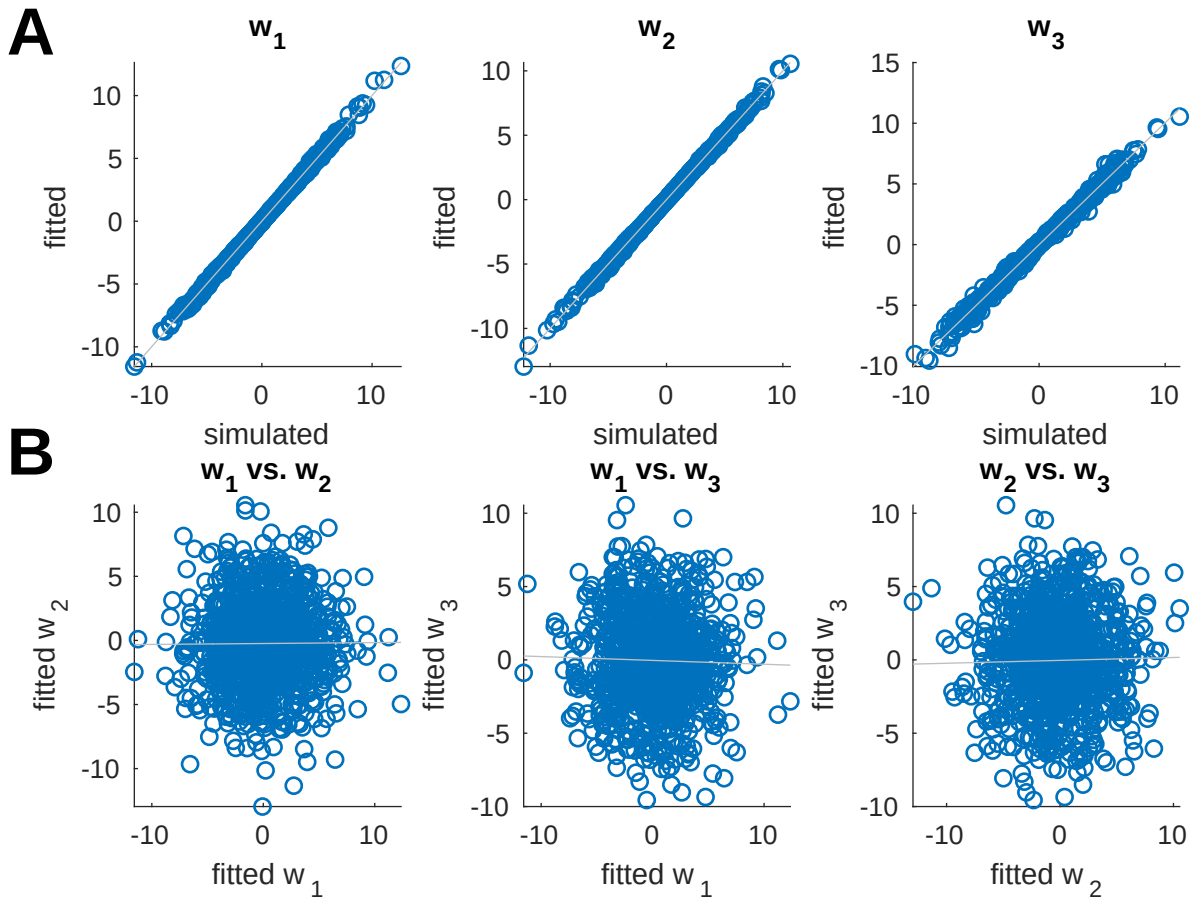

Figure S4. **Parameter recoverability**.

(A) Correlation between simulated and fitted parameters.

(B) Correlation between fitted parameters.

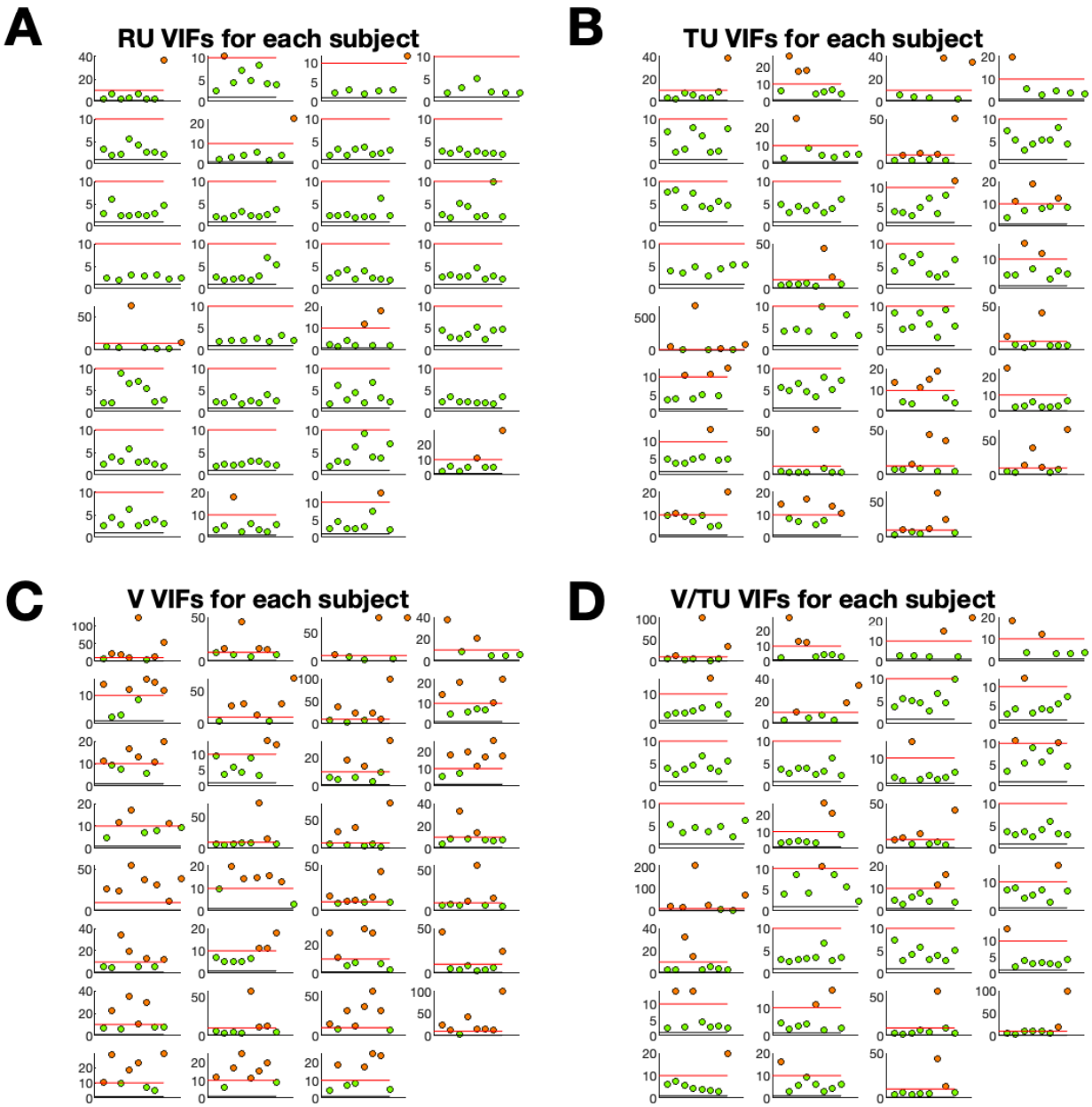

Figure S5. **Variance inflation factors** (VIFs) for parametric modulators of the trial onset regressor in the neural GLM. Each plot shows the VIFs for all runs of a given subject. Green circles correspond to runs with VIF <= 10, red circles correspond to runs with VIF > 10. A red horizontal line denotes the cutoff at 10.

(A) relative uncertainty (RU),

(B) total uncertainty (TU),

(C) value difference (V),

(D) value difference scaled by total uncertainty (V/TU).

Sign	**Brain region**	**BA**	**Extent**	***t*-value**	**MNI coord.**
Positive	Middle frontal gyrus (R)	47	141	4.503	34 48 -8
Negative	*Precuneus (R)	7	29494	-3.722	8 -72 48
	*Middle occipital gyrus (L)	18	29494	-9.441	-28 -90 6
	*Middle cingulate & paracingulate gyri (L)	32	2431	-7.037	-10 10 44
	*Precentral gyrus (L)	6	1428	-6.794	-38 -16 68
	*Midcingulate cortex (L)		597	-6.041	-16 -28 38
	*Lingual gyrus (L)		273	-5.425	-6 -44 2
	Thalamus (R)		135	-4.850	18 -30 8

Table S2. **RU (uncorrected)**. Brain regions in which the BOLD signal tracks RU (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 (L)	40	25410	9.342	-42 -40 52
	*IFG pars opercularis (L)	44	780	7.256	-42 4 28
	*Insula (L)	48	647	7.230	-28 20 4
	*Precentral gyrus (R)	44	687	6.943	48 8 30
	*Thalamus (L)		297	5.957	-4 -16 14
	*Insula (R)	47	312	5.304	36 18 0
	*Midbrain (L)		217	5.297	-8 -20 -12
	*Middle frontal gyrus (L)	46	671	5.202	-38 36 34
	Midcingulate cortex (L)	23	158	4.881	-2 -26 30
	Middle frontal gyrus (R)	46	119	3.968	38 34 34
Negative	*Superior frontal gyrus, dorsolateral (L)	10	1664	-7.101	-4 62 30
	*Precuneus (L)	30	446	-6.063	-8 -54 18
	Middle temporal gyrus (L)	21	120	-5.044	-56 -6 -18

Table S3. **TU (uncorrected)**. Brain regions in which the BOLD signal tracks TU (corresponding to Figure 4A). Notation and procedures as in Table S2. IFG, inferior frontal gyrus.

model	DF	AIC	BIC	LL	LR-stat	p-value
Hybrid model	3	7502.54	7524.01	-3748.27		
Augmented $\widehat{\mathrm{RU}}$ model	4	7500.34	7528.96	-3746.17	4.20	p = 0.040

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 48 -8). 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 (L)	-42 -40 52	7497.93	7526.56	-3744.97	6.61	$p = 0.0102$	$p = 0.0971$
IFG pars opercularis (L)	-42 4 28	7492.77	7521.40	-3742.39	11.76	$*p = 0.0006$	$*p = 0.0060$
Insula (L)	-28 20 4	7502.68	7531.30	-3747.34	1.86	$p = 0.1723$	$p = 0.8491$
Precentral gyrus (R)	48 8 30	7501.79	7530.41	-3746.89	2.75	$p = 0.0970$	$p = 0.6395$
Thalamus (L)	-4 -16 14	7502.35	7530.98	-3747.17	2.19	$p = 0.1388$	$p = 0.7756$
Insula (R)	36 18 0	7504.44	7533.06	-3748.22	0.10	$p = 0.7492$	$p = 1.0000$
Midbrain (L)	-8 -20 -12	7502.75	7531.38	-3747.38	1.79	$p = 0.1814$	$p = 0.8648$
Middle frontal gyrus (L)	-38 36 34	7492.70	7521.32	-3742.35	11.84	$*p = 0.0006$	$*p = 0.0058$
Midcingulate cortex (L)	-2 -26 30	7504.07	7532.69	-3748.03	0.47	$p = 0.4915$	$p = 0.9988$
Middle frontal gyrus (R)	38 34 34	7502.10	7530.73	-3747.05	2.44	$p = 0.1185$	$p = 0.7168$

Table S5. **$\widehat{\text{TU}}$ decoded from left DLPFC predicts choices.** Results from likelihood ratio test comparing the hybrid model (Eq. 4) with the augmented model which includes the neurally-decoded $\widehat{\text{TU}}$ (Eq. 12). Notation as in Table S4, with statistics included for the augmented $\widehat{\text{TU}}$ model only. * - significant after Bonferroni correction with adjusted $\alpha = 0.05/10 = 0.005$.

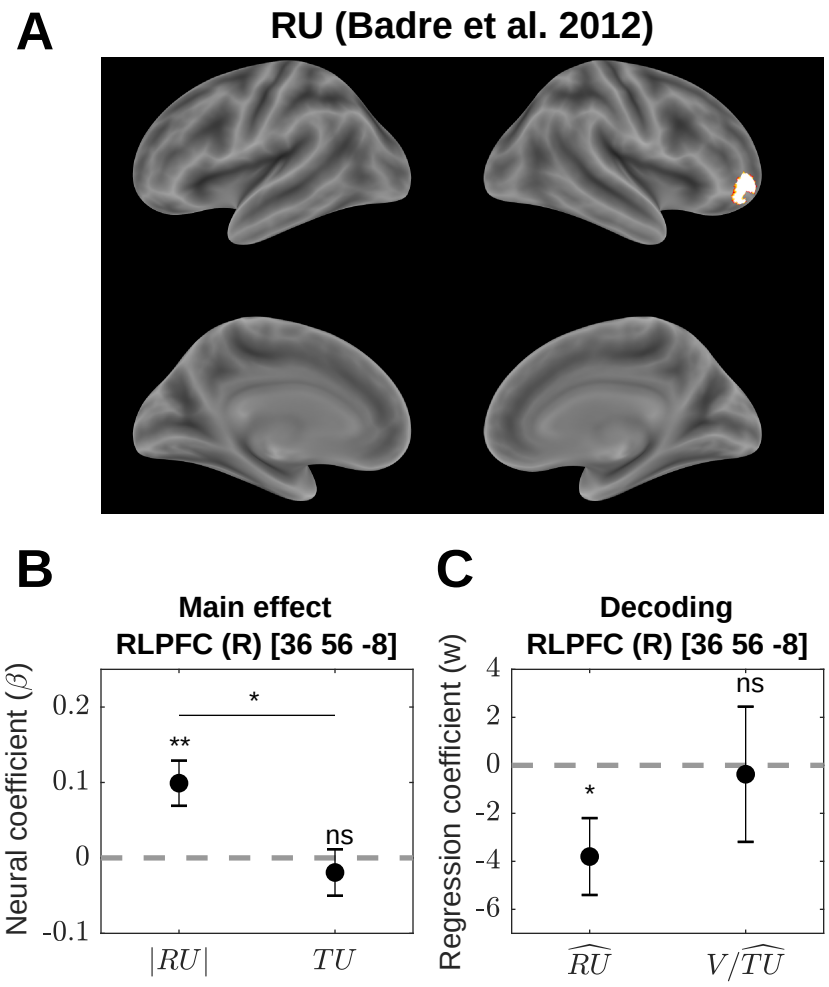

Figure S6. **Right RLPFC tracks RU but not TU.**

(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).

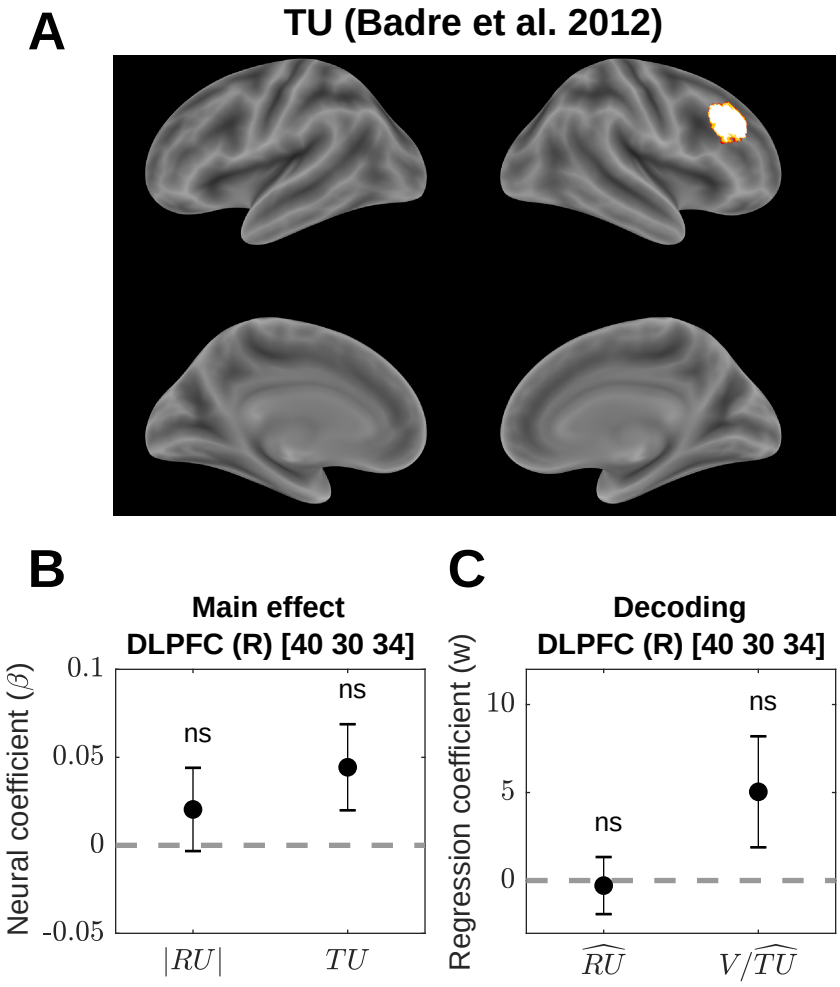

Figure S7. **Right DLPFC does not track TU.**

(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 4 but using the right DLPFC ROI from Badre et al. (2012).

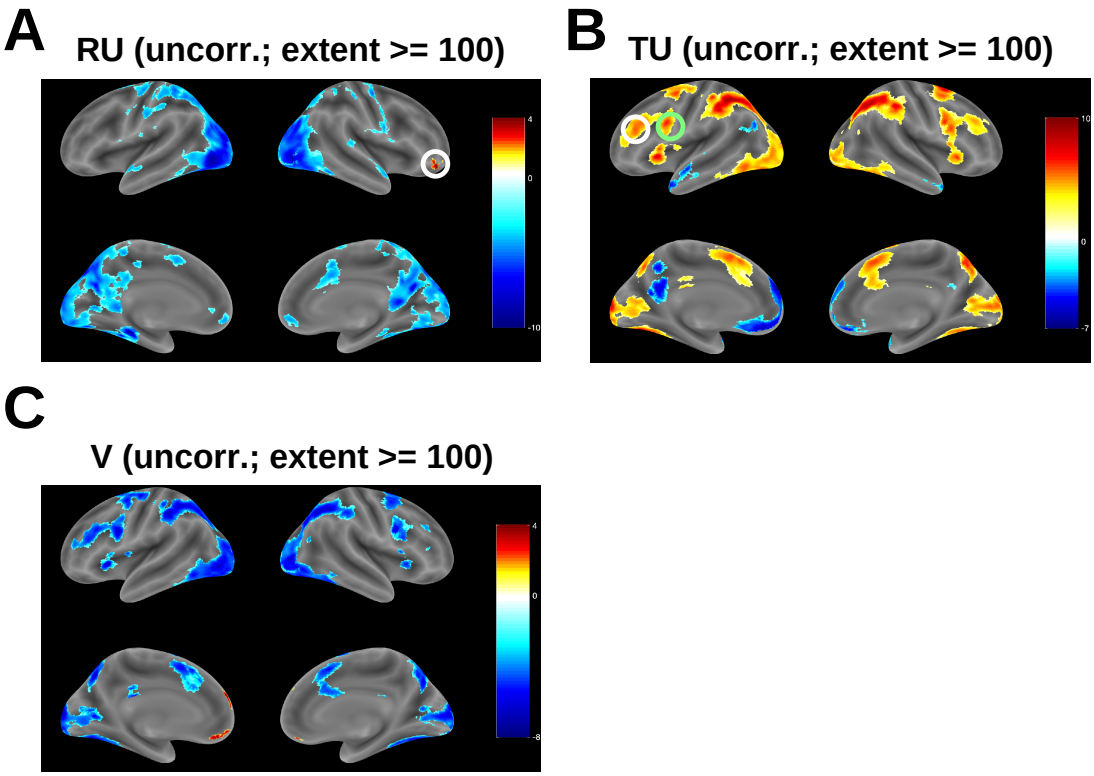

Figure S8. **Contrasts from GLMs with a single parametric modulator.**

(A) Whole-brain RU contrast when only RU was included as a parametric modulator. Compare with Figure 3A.

(B) Whole-brain TU contrast when only TU was included as a parametric modulator. Compare with Figure 4A.

(C) Whole-brain V contrast when only V was included as a parametric modulator.