Published in final edited form as:

Neuron. 2022 June 01; 110(11): 1869–1879.e5. doi:10.1016/j.neuron.2022.03.014.

The neurocomputational bases of explore-exploit decision making

Jeremy Hogeveen[1,2,9,*], **Teagan S. Mullins**[1,2], **John D. Romero**[1,2], **Elizabeth Eversole**[1,2], **Kimberly Rogge-Obando**[3], **Andrew R. Mayer**[1,4,5,6], **Vincent D. Costa**[7,8,*]

[1]Department of Psychology, University of New Mexico, Albuquerque NM 87131 USA.

[2]Psychology Clinical Neuroscience Center, University of New Mexico, Albuquerque NM 87131 USA.

[3]Department of Biomedical Engineering, Vanderbilt University, Nashville TN 37235 USA.

[4]Department of Psychiatry & Behavioral Sciences, University of New Mexico School of Medicine, Albuquerque NM 87131 USA.

[5]Department of Neurology, University of New Mexico School of Medicine, Albuquerque NM 87131 USA.

[6]The Mind Research Network/Lovelace Biomedical Research Institute, Pete & Nancy Domenici Hall, Albuquerque NM 87106 USA

[7]Department of Behavioral Neuroscience, Oregon Health and Science University, Portland OR 97239 USA.

[8]Division of Neuroscience, Oregon National Primate Research Center, Beaverton OR 97006USA.

Summary

Flexible decision making requires animals to forego immediate rewards (exploitation) and try novel choice options (exploration) to discover if they are preferable to familiar alternatives. Using the same task and a partially observable Markov decision process (POMDP) model to quantify the value of choices, we first determined that the computational basis for managing explore-exploit tradeoffs is conserved across monkeys and humans. We then used fMRI to identify where in the human brain the immediate value of exploitative choices and relative uncertainty about the value of exploratory choices were encoded. Consistent with prior neurophysiological evidence in monkeys, we observed divergent encoding of reward value and uncertainty in prefrontal and

[*]**Co-Corresponding Authors:** Jeremy Hogeveen, jhogeveen@unm.edu; Vincent D. Costa, costav@ohsu.edu.

Author Contributions: Conceptualization, J.H. and V.D.C.; Methodology, J.H. and V.D.C.; Software, J.H. and V.D.C.; Formal analysis, J.H. and V.D.C.; Investigation, J.H., T.S.M., J.D.R., E.E., K.R.-O., and V.D.C.; Resources, J.H., A.R.M., and V.D.C.; Writing, J.H., T.S.M., J.D.R., A.R.M., and V.D.C.; Project Administration, J.H., T.S.M., J.D.R., and V.D.C.; Funding Acquisition, J.H. and A.R.M.

[9]Lead Contact.

Declaration of Interests: The authors declare no competing interests.

parietal regions, including frontopolar cortex, and parallel encoding of these computations in motivational regions including the amygdala, ventral striatum, and orbitofrontal cortex. These results clarify the interplay between prefrontal and motivational circuits that supports adaptive explore-exploit decisions in humans and nonhuman primates.

eTOC Blurb:

How do humans and other animals make the decision to explore new options instead of exploiting familiar favorites? Hogeveen et al. find evidence for similar computations underlying explore-exploit decisions across humans and monkeys. Additionally, their study reveals a brainwide network comprising frontopolar, frontoparietal, frontostriatal, and mesocorticolimbic regions underlying explore-exploit decisions.

Introduction

The motivation to explore and acquire novel information shapes learning across the lifespan in many species. But exploration comes at the cost of exploiting familiar options whose immediate consequences are known. Managing this tradeoff is referred to as the explore-exploit dilemma. A barrier in identifying the neural bases of explore-exploit decision making is how to define a choice as exploratory. A common definition of an exploratory choice is the selection of an action that maximizes information rather than rewards. Defined in this way, it is clear that frontopolar and parietal regions play a role in the decision to explore new options and forego immediate rewards (Daw et al., 2006) and that lateral frontopolar cortex might drive exploration by tracking the relative uncertainty present in the choice environment (Badre et al., 2012; Cavanagh et al., 2012; Cockburn et al., 2021). An extension of this view is that explore-exploit decision making relies on prefrontal cortex to disrupt encoding of existing action and choice policies in motivational and sensorimotor circuits, while forming new decision policies (Choung et al., 2017; Daw et al., 2006; Domenech et al., 2020; Ebitz et al., 2018). But if a decision maker has sufficient knowledge of the environment and computational resources, exploration can be directed by value computations that, over the long-term, maximize gains or minimize losses (Averbeck, 2015; Wilson et al., 2021). In these cases, prefrontal cortex and motivational regions might work together to compute both the anticipated immediate and future value of choice options to determine when exploration of novel opportunities is advantageous (Costa and Averbeck, 2020; Costa et al., 2019; Tang et al., 2022; Wilson et al., 2020)

Neither of these perspectives are disjoint from one another. Decisions to explore that deviate from policies that maximize immediate rewards, likely take into account how choices are affected by uncertainty and the value of future rewards. But in humans it remains unknown if subdivisions of prefrontal cortex, particularly frontopolar cortex, explicitly encode computations derived from optimal decision strategies that define when it is advantageous to explore or exploit (Averbeck, 2015). It also is not clear if value encoding in prefrontal cortex during explore-exploit decisions occurs in parallel or in opposition to value encoding in motivational circuits, particularly during novelty-seeking. In nonhuman primates, neurons in dorsolateral prefrontal cortex (dlPFC) were found to encode *both* the immediate *and* latent future value of choice options to support adaptive explore-exploit

decisions (Tang et al., 2022) *prior to* making a choice. The same information is also encoded in a network of motivational brain regions comprising amygdala, ventral striatum, and orbitofrontal cortex *after* a choice is made and its outcome observed (Costa and Averbeck, 2020; Costa et al., 2019). This implies that prefrontal cortex and motivational brain regions cooperate with one another to manage explore-exploit tradeoffs. Consistent with this view, action encoding in prefrontal cortex is disrupted when monkeys decide to explore (Ebitz et al., 2018) and neurons in subthalamic and temporal lobe areas signal the availability of novel objects before they are viewed (Ogasawara et al., 2021). But at present, nonhuman primate studies are restricted to *a priori* targeted recording of one or a few brain regions at a time, which limits their anatomical scope. Moreover, frontopolar cortex can be challenging to access (Mitz et al., 2009), and has greatly expanded in humans relative to non-human primates (Mansouri et al., 2020).

No one has yet linked the insights simultaneously gained from human neuroimaging and neurophysiology experiments in macaques. In part, because it is unclear if humans and macaques use similar strategies to manage the explore-exploit dilemma. Here, we used the same multi-arm bandit reinforcement learning task to test humans and macaques (Costa and Averbeck, 2020; Costa et al., 2014, 2019; Wittmann et al., 2008) and used a partially observable Markov decision process model (POMDP;(Averbeck, 2015) to demonstrate that value computations underlying decisions to explore or exploit are conserved in primates. With this established, we then compared encoding of value computations associated with decisions to explore or exploit across the entire human brain and identified that prefrontal and motivational regions work together, rather than against one another in deciding when to explore. These results not only clarify the computational roles of prefrontal and motivational circuits in deciding when it is advantageous to explore, but also establish a translational bridge for future, complementary experiments in humans and nonhuman primates.

Results

Humans and Non-Human Primates Utilize Similar Computations to Manage Explore-Exploit Tradeoffs

Choice behavior.—We used a three-armed bandit task (Costa et al., 2019; Djamshidian et al., 2011; Wittmann et al., 2008) where the introduction of novel choice options was used to induce explore-exploit tradeoffs (Figure 1A–B). Novelty increases the value of exploring by increasing uncertainty about how choices affect future prospects for reward, but more importantly it makes the act of exploration explicit. Also, by referencing behavior to optimal decision strategies that let us derive the latent future value of making an exploratory or exploitative choice, can dissociate novelty-driven exploration from novelty seeking and or detection.

On each trial of the task, humans or monkeys viewed three choice options assigned different reward values. They had to learn the stimulus-outcome relationships by sampling each option (Figure 1A–B). However, the number of opportunities they had to learn about the value of each option was limited, as every so often one of the three options was randomly replaced with a novel option. Whenever a novel option was introduced uncertainty about its value was high, because humans or monkeys could not predict its assigned reward

probability. To reduce uncertainty and learn the value of the novel option, they had to explore it. But in doing so, they gave up the opportunity to exploit what they had learned about the two remaining options. We refer to the two remaining options as the best or worst alternative, based on how often their selection was rewarded in the past. See STAR Methods for a full description of the task and how choice behavior was analyzed.

Both humans and rhesus macaques preferred to explore the novel choice option instead of exploiting the value of the two remaining alternative options. Over the first few trials after a novel option was introduced, both humans and monkeys preferred to explore the novel option (Monkey: M=0.45, SEM=0.02; Human: M=0.46, SEM=0.02) rather than to exploit the best alternative option (Monkey: M=0.31, SEM=0.02, t=3.45, p=0.02, d=1.22; Human: M=0.35, SEM=0.02, t_{yuen}=2.96, p=0.007, η=0.55; Figure 1C–D). Both species also exploited what they had already learned by selecting the best alternative option more often than the worst alternative option (Monkey: M=0.24, SEM=0.01, t=2.76, p=0.03, d=0.97; Human: M=0.19, SEM_{human}=0.01, t_{yuen}=3.93, p<0.001, η=0.79; Figure 1C–D). Looking at decision making over time, there was a significant interaction between option type (i.e., novel, best alternative, and worst alternative) and trials since a novel stimulus was last inserted, on choice behavior in both humans (F=35.99, p<0.001) and monkeys (F=19.52, p<0.001). Across each primate species, this interaction was driven by opposing patterns of exploration and exploitation as a function of the number of trials since a novel stimulus was inserted. Specifically, the probability of selecting the novel stimulus decreased as the number of trials since a novel option was introduced increased (Monkey: b=−0.007, *95%CI*=−0.009 to −0.004, p<0.001; Human: b=−0.024, *95%CI*=−0.03 to −0.02, p<0.001; Figure 1E–F), whereas the probability of selecting the best alternative increased (Monkey: b=0.004, *95%CI*=0.002 to 0.006, p<0.001; Human: b=0.015, *95%*CI=0.01 to 0.02, p<0.001; Figure 1E–F). Collectively, behavioral performance indicated two common patterns across two different primate species: i) a preference to explore novel choice options to learn if they were more rewarding than previously chosen alternatives and ii) a sensitivity to the tradeoffs involved when choosing to explore or exploit to balance immediate and future opportunities to earn rewards.

Computational modeling of explore-exploit decisions.—Optimal strategies for managing explore-exploit trade-offs were estimated using a POMDP model, which accounts for uncertainty about future outcomes in estimating the value of taking particular actions. The utility of choosing an option is formally defined as the sum of two computations: immediate expected value (IEV; Figure 1G), which estimates the probability that choosing a particular option will immediately result in a gain, and future expected value (FEV), the discounted future gains that can be expected given what is learned after choosing a particular option. IEV is easy to compute by keeping track of how many times a particular option was chosen and how many times it resulted in a gain. Calculating the FEV is more demanding. It involves looking ahead a certain number of trials, simulating sequences of choices that could be made after choosing a particular action, and recursively estimating the number of future gains that would result from enacting each sequence to identify the best possible future outcome. Because subjects weren't limited to choosing the same option in future trials, the

FEV of a particular option mostly reflects the richness of the environment (is higher if the best available option has a high versus low IEV).

However, there are small differences in the FEV of each option that are critical in determining the value of exploring. These differences occur because the FEV of each option is tied to how often it has been chosen, which reflects uncertainty about its IEV. On each trial, the difference in the FEV of individual options from the average FEV of all three options quantifies the increase or decrease in future gains associated with choosing a particular option. We refer to this quantity as the exploration BONUS. In our task, when a novel option is introduced uncertainty about its IEV is high because it hasn't been sampled. Thus, the BONUS associated with choosing the novel option is highest when it is introduced and decreases over trials as the subject samples it and ascertains its value (Figure 1H, top). In parallel, the BONUS values associated with the best and worst alternative options are negative when a novel option is introduced (Figure 1H, bottom). This occurs because the subject has already sampled each option, lowering its FEV relative to the novel option. But, as the subject forgoes choosing alternative options to explore the novel option, the FEV and exploration bonuses for the alternatives increases. To summarize, choices with a positive BONUS value can be considered exploratory, whereas choices with a negative BONUS value can be considered exploitative.

We previously reported that the POMDP model outperforms several different choice heuristics in predicting when monkeys choose to explore or exploit (Costa et al., 2019), in particular a reinforcement learning model where novelty is assigned a fixed bonus in value (Costa et al., 2014; Kakade and Dayan, 2001; Wittmann et al., 2008). When we compared the ability of these two models to predict humans' choices the POMDP consistently fit better (M exceedance probability=83.78%; M BIC $_{\text{POMDP}}$= 402.87 vs. M BIC $_{\text{Novelty RL}}$= 421.28, M ΔBIC = 18.41), just as we had previously observed in rhesus macaques (Costa et al., 2019, 2020). To be clear, we do not think primates, humans or monkeys, mentally simulate all possible choice sequences to estimate exploration bonuses prior to each choice. But, the POMDP model does capture the fact that primates are utilizing uncertainty and the value of new information to inform their exploratory decisions. Also, the solutions used to generate value estimates in the POMDP model are based on basis function approximations that well characterize many of the processes carried out effortlessly by neural tissue (Poggio, 1990) and which may have developed on evolutionary timescales to allow for learning in dynamic environments (Kidd and Hayden, 2015).

To determine whether value estimates derived from the POMDP predicted choice in each primate species, we computed trial-by-trial estimates of the IEV and exploration BONUS for all three of the choice options on each trial. These normative estimates were independent from choice behavior. To determine the relative weighting of the IEV and exploration BONUS on each choice, we passed the POMDP model derived estimates for all choices through a softmax function, specifying two free parameters that scaled the IEV and exploration BONUS estimates. This procedure yielded trial-by-trial choice probabilities for each species' choices. If the POMDP well predicts choices, then the averaged choice probabilities for each option type should be correlated with the fraction of times a human or monkey chose that option. The correlations between POMDP model predictions and

observed choice data were positive and suggested a medium effect size across both humans (M=0.59, *95%CI*=0.47 to 0.70, t=10.5, p<0.001) and macaques (M=0.45, *95%CI*=0.30 to 0.61, t=7.43, p<0.001). Comparing the magnitude of these correlations across species, the POMDP did not differ in its ability to predict humans' or monkeys' choices (M_{diff}=0.13, *95%CI*=−0.03 to 0.30, t=1.64, p=0.09; Figure 1I). When we examined each of the free parameters in the softmax function, it was clear that IEV was a strong determinant of whether or not an option was chosen across species (Human: M=0.44, *95%CI*=0.26 to 0.63, t=4.96, p<0.001; Monkey: M=0.12, *95%CI*=0.01 to 0.22, t=2.82, p=0.037). However, exploration BONUS was not as strong as a predictor in humans as in monkeys (Human: M=0.02, *95%CI*=−0.12 to 0.16, t=0.26, p=0.80; Monkey: M=0.08, *95%CI*=−0.002 to 0.16, t=2.52, p=0.05). Inclusion of the BONUS parameter, however, did improve our overall ability to predict human participants' choices compared to when the parameter was excluded from the model (χ^2_{37}=116.68, p<0.001; M BIC_{+BONUS}=421.28 vs. M BIC_{-BONUS}=595.46, M ΔBIC=174.17). Perhaps novelty was more salient to the monkeys because more trials had elapsed between the introduction of novel options relative to the human paradigm (i.e. longer time horizon; Wilson et al., 2021), or because individual differences in novelty-driven exploration are more apparent when secondary versus primary reinforcement is used (e.g. we only observed one monkey that was not prone to exploring novel options). But it is important to point out that the free parameters for the IEV and exploration BONUS were fit together and more relevant than whether individual parameters differed from zero, is whether there was a consistent relationship between the two parameters across humans and non-human primates. This would confirm that despite individual differences in subjects' tendencies to explore or exploit, relative weighting of the uncertainty surrounding these decisions informed how each species chose to manage explore-exploit tradeoffs. In both humans and monkeys, we observed comparable negative correlations between IEV and exploration BONUS parameters across species (Human: *rho*=−0.48, *95% High Density Interval (HDI)*=−0.73 to −0.19; Monkey: *rho*=−0.67, *95%HDI*=−0.97 to −0.04; Figure 1J; Supplementary Figure 2).

Neurocomputational Bases of Novelty-Driven Exploration in Humans

Encoding of the bonus in future value that predicts exploratory choices.—We determined whether—as in nonhuman primates (Costa and Averbeck, 2020; Costa et al., 2019)—the human brain encodes the potential increase in future value that can be acquired through exploration. To test this, we modeled trial-to-trial variance in choice-evoked brain activity as a function of exploration BONUS value estimates derived from the POMDP model (STAR Methods). For each subject, first-level model coefficients, that related trial-by-trial changes in BOLD activity to the exploration BONUS for chosen options, were extracted from 370 ROIs spanning the entire cortex, medial temporal lobes, and dorsal and ventral striatum. These parameter estimates were passed to a group-level Bayesian Multi-Level Model (BMLM; (Chen et al., 2019) to determine which ROIs encoded the exploration BONUS estimates from the POMDP model. BMLM approaches are gaining in popularity for both functional activation and connectivity-based fMRI studies (Cosme et al., 2021; Lima Portugal et al., 2020; Limbachia et al., 2021; Yin et al., 2019) because they circumvent multiple comparison problems commonly encountered with univariate approaches and produce highly overlapping results compared to conventional voxel-level

modeling with improved modeling efficiency (cf., (Limbachia et al., 2021); Supplementary Figure 1).

When participants chose an option there was widespread encoding of its associated exploration BONUS value. In most brain regions that encoded the exploration BONUS, neural activity was aligned with the sign of the POMDP derived model estimate. Positive encoding was observed in dorsolateral prefrontal cortex (e.g. posterior area 9/46), ventrolateral prefrontal cortex (e.g. inferior frontal sulcus), ventromedial prefrontal (e.g. area 10v), orbitofrontal cortex, rostral anterior cingulate cortex (e.g. area 32), posterior parietal regions (e.g. lateral intraparietal area), inferior temporal lobe regions (e.g. area TE), and visual areas (e.g. fusiform face complex; Supplementary Table 1; Figure 2). We only found two brain regions in which neural activity was inversely aligned with the sign of the exploration BONUS. Negative encoding was found in lateral frontopolar cortex (e.g. area p10p, anterior 9/46), and posterior cingulate cortex (e.g. area 31; Supplementary Table 1; Figure 2). It is often argued that differences in neuronal sources can determine whether BOLD responses are positive or negative: Whereas an association between positive BOLD and excitatory neuronal activity is well-established (Logothetis et al., 2001), neuronal inhibition in deep cortical layers may lead to vasoconstriction and a local increase in deoxyhemoglobin, leading to negative BOLD signal (Shmuel et al., 2006). Therefore, regardless of whether a region exhibited positive or negative deflections in activity as a function of exploration BONUS—the more important point is that all of the identified regions likely play a computational role in directed exploration in some manner.

Overlapped and distinct encoding of perceptual novelty and exploration value.—There is an important distinction between exploration driven by perceptual novelty–either due to surprise or habituation–and exploration motivated by an explicit desire to gain information, reduce uncertainty, and maximize future rewards (Averbeck, 2015). Recognizing this distinction, we included regressors in the first-level fMRI model that separately accounted for i) variance explained by the exploration BONUS associated with choices, and ii) the number of trials that had elapsed since a novel option was introduced. Because participants did not always choose the novel option and the sign of the exploration bonus associated with each option switched from positive to negative based on how often they were sampled, these two regressors were only weakly correlated. This allowed us to control for variance in neural activity driven by perceptual novelty when assessing encoding of POMDP derived valuations, since whether or not the novel option was chosen it was always present on the screen at the time a choice was executed. But it also allowed us to detect brain regions which encoded perceptual novelty by passing first-level coefficients for the perceptual novelty regressor to a group-level BMLM.

Assessing the potential conjunction of these effects (i.e., perceptual novelty and exploration value) is of interest because although prior neurophysiology and neuroimaging studies have identified a large set of brain regions involved in novelty processing, encoding of perceptual novelty is weak relative to encoding of POMDP derived value signals in the few brain regions that have been examined in macaques (Costa and Averbeck, 2020; Costa et al., 2019; Tang et al., 2022). A conjunction map of perceptual novelty and exploration BONUS encoding revealed a diffuse set of frontoparietal network and subcortical regions which

encoded both factors (Supplementary Table 2; pink in Figure 4). But we also found through this analysis that there were multiple regions which exclusively encoded either perceptual novelty or the value of exploring. Distinct encoding of the number of trials that had elapsed since the introduction of a novel choice option was found in temporopolar regions, inferior temporal cortex, and components of the brain's 'salience network' (anterior insula and dorsal anterior cingulate cortex; Supplementary Table 2; orange in Figure 4). Whereas distinct encoding of the relative gain or loss in future value associated with exploring or exploiting was found in frontopolar cortex, ventromedial prefrontal cortex, rostral anterior cingulate cortex, and nucleus accumbens (Supplementary Table 2; Figure 4, blue in Figure 4).

Encoding of immediate reward value that predicts exploitative choices.—We modeled trial-to-trial variance in choice-evoked brain activity as participants learned the IEV associated with choosing a particular option. Neural activity at the time of choice was positively related to the IEV of the chosen option in bilateral ventromedial prefrontal cortex (e.g. area 10v), posterior cingulate regions (e.g. area 31), bilateral somatomotor regions (e.g. area 4), anterior temporal regions (e.g. area TE), and visual cortex. Conversely, neural activity within nodes of frontoparietal and cingulo-opercular brain networks were negatively correlated with the IEV of the chosen option, including dorsolateral prefrontal cortex (e.g. posterior area 9/46), dorsomedial prefrontal and anterior cingulate regions (e.g. medial area 8; area 32), the lateral intraparietal area, and anterior insula (Supplementary Table 3; Figure 2B).

Dissociable encoding of explore-exploit value computations in dorsal versus ventral circuitry.—We examined whether brain regions we identified as encoding both BONUS and IEV showed similar or opposite patterns of encoding across these two models. First, we found that lateral frontopolar cortex more strongly encoded the potential gain or loss in future value associated with choosing to explore or exploit, whereas dorsal ACC and a cluster of somatomotor regions more strongly encoded the immediate likelihood that a choice would be rewarded. All of the other brain regions we identified through our BMLM analyses encoded both decision variables and clustered together based on whether they encoded the IEV and exploration BONUS in a similar or opposing manner. Brain regions wherein activity at the time of choice encoded the IEV and exploration BONUS of chosen option in an opposing manner included dlPFC, the dorsal subdivision of the lateral intraparietal area, and dorsal anterior cingulate cortex (Figure 2C). Whereas activity similarly encoded the IEV and exploration BONUS in ventromedial prefrontal cortex, orbitofrontal cortex, rostral anterior cingulate (Figure 2C), nucleus accumbens, and amygdala (Figure 3B). These distinct encoding patterns between dorsal frontostriatal and ventral mesocorticolimbic circuitry suggests these dorsal and ventral circuits could play distinct roles in deciding when to explore or exploit (Averbeck & Murray, 2019; Tang et al., 2022).

Discussion

Humans and rhesus macaques both attempt to resolve uncertainty about how their current choices will affect future outcomes when confronted with explore-exploit tradeoffs. By

modeling humans' and monkeys' choices in a unifying computational framework we were able to broaden insights into the neural correlates of explore-exploit decisions gleaned from neurophysiology experiments in monkeys, and identify broad networks of prefrontal and motivational brain regions in humans that contribute to this complex form of model-based reinforcement learning. Similar to recent monkey neurophysiological data (Costa and Averbeck, 2020; Costa et al., 2019; Tang et al., 2022), we found distributed, but distinctly patterned encoding of POMDP derived value computations necessary for successfully managing explore-explore tradeoffs. Distinct encoding patterns were observed in key subdivisions of human prefrontal cortex, including frontopolar cortex, as well as in motivational brain regions such as the amygdala, ventral striatum, and orbitofrontal cortex. This suggests that homologous neural circuits aid humans and nonhuman primates in solving the explore-exploit dilemma. Moreover, by taking advantage of our ability to map these value computations across the entire human brain our data suggest that there is a much more dynamic interplay between frontopolar cortex, frontoparietal, frontostriatal, and mesocorticolimbic circuitry during explore-exploit decision making than is currently hypothesized (Daw et al., 2006; Mansouri et al., 2020; Wilson et al., 2021).

The observation that neural activity in frontopolar cortex encoded the small differences in the relative future value that signaled when exploration is advantageous helps to resolve the computational role of this subdivision of prefrontal cortex unique to primates (Wise, 2008), which has been routinely implicated in explore-exploit decision making (e.g. Daw et al., 2006; Mansouri et al., 2017; Zajkowski et al., 2017). Enhanced deactivation of multiple subdivisions of frontopolar cortex (area p10p and anterior area 9/46) was observed when participants selected options with a high exploration BONUS value (i.e. novel options), and this result was orthogonal to encoding perceptual novelty and its habituation. Notably, in line with human lesion-symptom mapping studies showing aberrations in value-based choice in patients with damage to ventromedial prefrontal and rostral anterior cingulate cortex (Hogeveen et al., 2017; Kovach et al., 2012; Reber et al., 2017), ventromedial frontopolar cortex (area 10v) encoded the anticipated decision value of the chosen option—evidenced by parallel encoding of both exploration BONUS and IEV of choices. In contrast, posterolateral frontopolar regions encoded exploration BONUS, but did not demonstrate strong encoding of IEV or perceptual novelty. Disruptive brain stimulation applied to posterolateral frontopolar regions selectively impairs directed, but not random exploration (Zajkowski et al., 2017). These data suggest that whereas ventromedial frontopolar cortex encodes the decision value of a choice (relevant to both exploration or exploitation), posterolateral frontopolar cortex is more strongly involved in the decision to explore new options and maximize future opportunities to earn reward.

Our finding that posterolateral frontopolar cortex (area p10p) encodes the bonus in value associated with exploring versus exploiting helps to resolve competing hypotheses about its computational role. Posterolateral frontopolar cortex has consistently been implicated in exploration in humans (Badre et al., 2012; Boorman et al., 2009; Daw et al., 2006), but interpretations about its computational role are less consistent. One view suggests this frontal subdivision encodes the value of information that resolves uncertainty when participants choose to explore (Badre et al., 2012). A different perspective is that this same region is involved in top-down inhibition of action selection circuitry to enable

switching from exploitative to exploratory actions (Daw et al., 2006). These hypotheses are not mutually exclusive, and our data provide some support for both notions. First, in our approach the POMDP model assumes that an agent is using uncertainty about the future to guide exploration. Therefore, our results showing that posterolateral frontopolar cortex is increasingly deactivated as the bonus associated with exploring uncertain options increases, are well aligned with existing studies showing that activity in area p10p scales with decision uncertainty (Badre et al., 2012). Additionally, because the exploration BONUS associated with a choice was negatively encoded in area p10p and positively encoded in frontoparietal networks, local white matter tracts between area p10p and more posterior prefrontal regions (Baker et al., 2018) may enable predictions about uncertainty and future value to bias downstream recruitment of frontoparietal action selection circuits. Causal circuit manipulation studies contrasting the computational roles of different regions in the current task across primate species are warranted to test these ideas.

Several regions of inferior temporal cortex, temporopolar regions, anterior insula, and dorsal anterior cingulate cortex demonstrated unique encoding of perceptual novelty and its habituation, related to when a novel stimulus was introduced as a choice option and repeatedly viewed across trials despite it being chosen or not. These same regions did not encode the exploration BONUS associated with choosing novel or alternative options. Primates respond preferentially to novel events, and one of the mechanisms for this is thought to be adaptations in the firing rate of inferior temporal cortex neurons to the relative familiarity of visually-presented objects (Desimone, 1992; Jaegle et al., 2019; Rodman, 1994). Since BONUS estimates for novel options decreased non-linearly with repeated sampling over trials, it is also possible that some of the shared variance between BONUS and novelty in visual cortices was associated with a non-linear decay of perceptual novelty responses in these regions (Sutton and Barto, 1990). Beyond this novelty-orienting response, recent recordings from nonhuman primates suggest that anterior inferior temporal neurons play a critical role in the control of subsequent novelty-seeking behaviors (Ogasawara et al., 2021). Finally, insula and dorsal anterior cingulate have been argued to represent nodes within the brain's 'salience network', which among other functions, is thought to play a role in exogenous attentional capture by highly salient sensory events (Corbetta et al., 2008; Uddin et al., 2017). Therefore, isolating variance in BOLD activation that was associated with the degree of novelty of options within the choice set revealed brain maps that are in direct agreement with prior systems neuroscience research on the neuronal mechanisms of novelty-driven attentional orienting and novelty-seeking behaviors.

Lastly, we observed dissociations between explore-exploit computations in corticostriatal brain networks. Specifically, in the frontoparietal network and dorsal striatum BOLD activity was increased as a function of exploration BONUS, suggesting increased recruitment of these regions when individuals decided to explore novel options and resolve uncertainty about their future value. This would complement the known role of frontoparietal regions in matching behavior (Sugrue et al., 2004), foraging (Genovesio, Wise, and Passingham, 2014), and information sampling (Furl and Averbeck, 2011; Costa and Averbeck, 2015). Several frontoparietal regions were also negatively associated with IEV, indicating deactivation during the exploitation of familiar rewards. One possibility is that participants rapidly form an internal model of the general structure of the task, which

includes an expectation about how frequently familiar options would be replaced with novel options. This acquired model of the task would enable the subject to infer that novel stimuli provide a potential increase in future value relative to familiar options (in accordance with the exploration BONUS parameter from the POMDP). Conversely, after participants have had the opportunity to repeatedly sample novel choice options and learn whether they are better or worse than familiar alternatives (i.e., when all options are familiar, and their IEV is differentiated), participants may shift from use of circuits enabling model-based control to parallel systems implicated in model-free stimulus-outcome learning. In this view, when exploration BONUS is high during explore-exploit decision making, brain regions involved in model-based control (e.g. dorsolateral prefrontal cortex, lateral intraparietal cortex, and dorsomedial striatum; (Averbeck and O'Doherty, 2022; Gläscher et al., 2010; Liljeholm and O'Doherty, 2012; Smittenaar et al., 2013)) demonstrate greater activation due to the increased engagement of model-based predictions about the potential future value of exploration. In contrast, when IEV is high, the anticipated value of the chosen option is well-learned and model-free control can operate, associated with less intensive neural computations in model-based control circuits (cf., (Otto et al., 2013; Smittenaar et al., 2013)). These notions would also accord with the view that dorsal frontostriatal circuits play a role in Bayesian state inference and goal-directed behavioral control more broadly (Averbeck and Murray, 2020; Bartolo and Averbeck, 2020). Additionally, our finding of enhanced lateral intraparietal activity when exploration BONUS is high, and decreased intraparietal activity when IEV is high, is compatible with existing studies dissociating novelty and reward anticipation in this region in nonhuman primates (Foley et al., 2014).

But there still remains an outstanding question as to how deactivation in frontopolar cortex corresponds to activation in model-based control regions when uncertainty motivates decisions to explore. One possibility is that model-based deactivation in posterolateral frontopolar (p10p) is related to changes in the activity of inhibitory interneurons in this area, given that in macaques frontopolar cortex neurons are task engaged throughout each trial but only encode goals at the time of feedback (Tsujimoto et al., 2011). This also fits with the observation that exploration BONUS encoding is delayed in anterior vs. posterior dorsolateral prefrontal cortex (Tang et al., 2022) indicative of feedforward processing. Future cell-type specific neurophysiology and neuromodulatory studies examining frontopolar, frontoparietal, and frontostriatal networks in nonhuman primates are needed to better understand the consistent engagement of this region during exploratory decision making.

Overall, we report evidence that humans and monkeys perform similar neural computations when exploring novel stimuli in lieu of exploiting familiar rewards. Across primate species we observed a novelty-driven exploration decision bias in the immediate wake of encountering new choice opportunities, alongside an increased tendency to exploit the best available option relative to the worst available alternative. These decision tendencies were well-predicted by value computations derived from a POMDP model of explore-exploit decision making. Therefore, the motivation to explore novel options and maximize future value when confronted with the explore-exploit dilemma represents an exciting avenue for cross-species primate research, providing a new bench-to-bedside pipeline for interventions for pathological reward processing or novelty sensitivity in clinical populations.

STAR Methods

Resource Availability

Lead contact.—Requests for further information should be directed to and will be fulfilled by the lead contact, Jeremy Hogeveen (jhogeveen@unm.edu).

Materials availability.—This study did not generate new unique reagents.

Data and code availability.—All original code from this paper has been shared via GitHub, and a DOI release created via Zenodo. The code is publicly available as of the date of publication, and the DOI is listed in the Key Resources Table.

Human Participants

47 adult participants from the Albuquerque, New Mexico community were enrolled in the current study, which was approved through the University of New Mexico Office of the Institutional Review Board. Unfortunately, N=5 of 47 enrolled participants were unable to complete their scheduled study visit due to mandated COVID19 pandemic lockdown restrictions in Spring, 2020. Additionally, N=1 participant was removed from study analyses due to extreme non-normative behavior relative to the computational model (3.34 residual standard deviations from the group line of best fit between the BONUS and IEV coefficients), N=1 participant was removed due to excessive head motion during fMRI (>3 standard deviations above mean framewise displacement across all task fMRI runs), N=2 participants were removed due to insufficient responding during the task (N=144 responses during the task, −4.11 standard deviations below mean across the sample), and N=1 participant was removed due to an incorrect key configuration on the task fMRI button box. Therefore, the final sample in the current study comprised N=37 participants.

This final sample included 24 female and 13 male participants (M_{age}=26.6 years; SD=7.24 years). 12 participants were of Hispanic or Latino ethnicity, 20 were Not Hispanic or Latino, and five did not choose to report their ethnicity. The final sample included two mixed race participants (one Black or African American and American Indian/Alaska Native, one American Indian/Alaska Native and White), three Asian participants, one Black or African American participant, 26 White participants, and five did not choose to report their race.

Monkey Subjects

Eight adult male rhesus macaques (*Macaca mulatta*) served as subjects. Their ages and weights at the start of training ranged between 6–8 years and 7.2–9.3 kg. Animals were pair housed when possible, had access to food 24 hours a day, were kept on a 12-h light-dark cycle, and tested during the light portion of the day. On testing days, the monkeys earned their fluid through performance on the task, whereas on non-testing days the animals were given free access to water. All procedures were reviewed and approved by the NIMH Animal Care and Use Committee. Behavioral data from a subset of the same monkeys also appears in x(Costa et al., 2019).

Novelty-Bandit Task

Human version.—Participants made speeded (≤2 seconds) manual responses between three neutral images taken from the International Affective Pictures System (IAPS(Bradley and Lang, 2020)). Stimuli were presented using EPrime Version 3 (Psychology Software Tools, Sharpsburg, PA), with responses recorded using a MIND Input Device (https://www.mrn.org/collaborate/mind-input-device). Images were randomly assigned an *a priori* low (p=0.2), medium (p=0.5), or high (p=0.8) reward probability. Every 5–12 trials (M≈6) a 'novel insertion' took place, wherein one familiar image in the current set was replaced by one novel image to create a new set that would be presented for the proceeding 5–12 trials. Novel images were randomly assigned a low, medium, or high reward probability, with the caveat that all 3 images could not have the same assigned reward probability in the new set. Participants completed 224 trials containing 32 novel stimulus insertions, divided evenly into 4*≈7-minute fMRI runs. Participants made confidence judgments (low, medium, or high confidence) after each decision, but these data are not relevant to the current manuscript. Image location was randomized on each trial, and participants received either reward (green '+1') or nonreward (red '0') feedback after each decision. Notably, the fixation cross jittered durations were optimized to maximize efficiency for deconvolving the hemodynamic response at the time of choice, choice option were presented centrally to minimize saccade-related BOLD activity, manual responses indicated choice, and feedback was a symbolic cue (+1 versus 0; Figure 1A).

Monkey version.—Subjects performed a saccade-based version of the same task performed by humans. Stimulus presentation and behavioral assessment were controlled via Monkeylogic (Hwang et al., 2019), and eye movements were sampled at 400 frames per second, 1000 Hz using an Arrington Viewpoint eye tracker (Arrington Research, Scottsdale, AZ). Each session began with three naturalistic scenes randomly assigned an *a priori* low, medium, or high probability of being paired with an apple juice reward via a precise liquid-delivery device (Mitz, 2005). Every 8–30 trials (M≈19 trials) a novel insertion took place, and the novel image was assigned a low, medium, or high reward probability, with the caveat that all three images in the new set could not have the same reward likelihood. Monkeys completed 650 trials per session, with each trial beginning with a central fixation (250–750ms), followed by the presentation of three images in the periphery, and a saccade to and fixation on one of the targets for 500ms). Subjects performed a saccade-based version of the same task performed by humans. In the monkey experiment, the trial initiated when the animal maintained fixation for 0.5–0.75s, options were presented peripherally to minimize the influence of stimulus position on choice, saccades indicated choice, and feedback was a juice reward (Figure 1A). Stimulus presentation and behavioral assessment were controlled via Monkeylogic (Hwang et al., 2019), and eye movements were sampled at 400 frames per second, 1000 Hz using an Arrington Viewpoint eye tracker (Arrington Research, Scottsdale, AZ).

Computing choice probabilities.—To assay discrete explore-exploit decisions in humans and rhesus macaques, we computed the likelihood that each subject would select the novel option, the best alternative option (i.e., the familiar option that had most often resulted in prior reward), and the worst alternative option (i.e., the familiar option that had least often

resulted in prior reward) during the early trials after a novel stimulus was inserted. These "early trials" were defined as the first N=2 trials post-insertion for humans. Relative to the modal interval between insertion trials (M≈6 for humans, M≈19 for monkeys), this was translated to N=6 post-novel trials for the monkeys.

Additionally, to model decision making over time, we computed the probability of selecting the novel, best, or worst alternatives as a function of trials post-novel stimulus insertion. To be clear, we computed the probability that each participant would select each option on 0 trials post-novel, 1 trial post-novel, 2 trials post-novel, etcetera. Notably, there were low trial numbers in the human dataset for >6 trials post-novel, and >19 trials post-novel for macaques. Therefore, the choice frequencies on these late run trials were summed together (e.g., for human data the probability of choosing the best available option on trial post-novel=6 in our model actually reflects the probability of choosing the best alternative option on trials post-novel≥6).

Assumptions of the General Linear Model.—We evaluated whether or not choice probability data met the assumptions of normality and homoscedasticity. We observed significant evidence for violations to the assumption of normality within human (probability of choosing the novel stimulus: W=0.97, p<0.001; probability of choosing the best alternative: W=0.98, p=0.007) but not monkey (probability of choosing the novel stimulus: W=0.99, p>0.2; probability of choosing the best alternative: W=0.99, p>0.1) subjects. We also observed significant evidence for a violation of the assumption of homoscedasticity for the model fit contrast between human and monkey subjects (F=4.55, p=0.044). Therefore, robust inferential tests were used for human behavioral analyses, and cross-species comparisons, but conventional tests were used within the monkey subjects. Regarding the robust inferential tests used for human choice data, for comparisons between two conditional means we used Yuen's modified t-test for trimmed means [t_{yuen}; (Yuen, 1974)]. For our analysis of decision probabilities over trials, we computed robust linear mixed effects models that down-weight observations with large residuals and reduce their impact on model estimates (Koller, 2016). Importantly, none of the conclusions derived from these robust inferential tests would be reversed by, instead, using conventional t-tests or linear mixed model estimators.

Partially Observable Markov Decision Process Model

The task was modeled using a partially observable Markov decision process (POMDP) model. In the POMDP model, utilities are defined by the information state, which is a hidden variable that can be inferred based on observed choices and outcomes, across all three options. The set of possible next information states that can be reached is determined by whether or not the current choice is rewarded and whether one of the three options is replaced with a novel option. Thus, each choice leads to 21 unique subsequent states and each choice after that to another 21 unique states. Looking out over a particular time horizon the future information states that can be reached after the current choice can be represented by a binary tree. The information state reflects a product space across the three binary trees formed for each option. Transitions through the information state space occur after each choice and its associated outcome, and they correspond to belief updates for the POMDP.

Because the state space quickly becomes intractable over relevant time horizons, we used approximation methods to fit an infinite horizon, discrete state, discounted POMDP using B-spline basis functions to estimate the value of information states (Averbeck, 2015). The utility, u, of a state, s, at time t is:

$$u_t(s_t) = max_{a \in A_{s_t}} \left\{ r(s,a) + \gamma \sum_{j \in s} p(j \mid s_t, a) u_{t+1}(j) \right\} \quad (1)$$

Where A_{s_t} is the set of available actions in state s at time t, $r(s_t, a)$ is the reward that will be obtained in state, s, at time, t, if action, a, is taken. $\sum_{j \in s}$ represents the summation of all possible subsequent states at, $t + 1$, or the expected future utility taken across the transition probability distribution, $p(j|s_t, a)$. Transition probability refers to the probability to transitioning into each future state, j, from the current state, s_t, if the subject takes action, a. Gamma represents a discounting term set at 0.9. The terms inside the curly braces represent the utility for each available option, $Q(s_t, a) = r(s, a) + \gamma \sum_{j \in s} p(j|s_t, a) u_{t+1}(j)$. The immediate expected value (IEV) refers to the first term in the utility function, $IEV = r(s, a)$, while the future expected value (FEV) is the second term, $FEV = \gamma \sum_{j \in s} p(j|s_t, a) u_{t+1}(j)$. The exploration BONUS is the FEV of a given action, a, relative to the average FEV of all available options, $BONUS(a) = FEV(a) - (\sum_{j=1:3} FEV(j))/3$.

We used a value iteration algorithm to fit utilities (Puterman, 1994), in which the vector of utilities across states, v^0, was initialized to random values at iteration, $n = 0$, and then updated by computing:

$$v^{n+1} = max_{a \in A_{s_t}} \left\{ r(s,a) + \gamma \sum_{j \in s} p(j \mid s_t, a) v^n(j) \right\} \quad (2)$$

Following each iteration the change in value was calculated as, $\Delta v = v^{n+1} - v^n$, and examined either $||\Delta v|| < \in$ or $span(\Delta v) < \in$. The span is defined as $span(v) = max_{s \in S}(s) - min_{s \in S}(s)$.

The state space was intractable over relevant time horizons, therefore approximation methods involving b-spline basis functions and utility approximation were used (Friedman et al., 2001), with:

$$\hat{v}(s) = \sum_{i=1}^{m} a_i \Phi_i(s) \quad (3)$$

We used fixed basis functions so we could calculate the basis coefficients, a_i, using least squares techniques. We assembled a matrix, $\phi_{i,j} = \phi_i(s_j)$, which contained the values of the basis functions for specific states, s_j. We then calculated a projection matrix:

$$H = \phi(\phi'\phi)^{-1}\phi' \quad (4)$$

And calculated the approximation:

$$\hat{\boldsymbol{v}} = H\boldsymbol{v} \quad (5)$$

Where bold indicates a vector over states, or the sampled states at which we computed the approximation. When using the approximation in the value iteration algorithm, we first compute the approximation, $\hat{v}$. This approximation was then substituted for the v in equation 2 (i.e., $max_{a \in A_{s_t}} \left[r(s,a) + \gamma \sum_{j \in s} p(j \mid s_t, a) \hat{v}^n(j) \right.$. Approximations to new values are then calculated as, $\hat{\boldsymbol{v}}^{\boldsymbol{n+1}} = H\boldsymbol{v}^{\boldsymbol{n+1}}$, until convergence.

State space definition.—The novelty task is a three-armed bandit task. The options are rewarded with different probabilities, but the amount of reward is always 1. The reward probabilities for each bandit are stationary while that option is available. On each trial there is a 5% chance that one of the bandit options will be replaced with a new option. The underlying model is a discrete MDP. The state space is the number of times each option has been chosen, and the number of times it has been rewarded, $s_t = R_1, R_1, R_2, C_2, R_3, C_3$.

This state space was approximated using a continuous approximation sampled discretely. The immediate reward estimate is given by the maximum *a-posteriori* estimate, $r(s_t, a = i) = \frac{r_i + 1}{c_i + 2}$. The set of possible next states, $s_t + 1$, is given by the chosen target, whether or not it is rewarded, and whether one of the options is replaced with a novel option (Averbeck et al., 2013). Thus, each state leads to 21 unique subsequent states. We define $q_i = r_t(s_t, a = i)$, and $p_{switch} = 0.05$, as the probability of a novel stimulus substitution. The transition to a subsequent state without a novel choice substitution and no reward is given by:

$$p_t(\ldots, C_i + \boldsymbol{1}, R_i, \ldots \mid s_t = [\ldots, C_i, R_i, \ldots], a = choose\ i) = (\boldsymbol{1} - q_i)(\boldsymbol{1} - p_{switch}) \quad (6)$$

And for reward by:

$$p_t(\ldots, C_i + \boldsymbol{1}, R_i + \boldsymbol{1}, \ldots \mid s_t = [\ldots, C_i, R_i, \ldots], a = choose\ i) = q_i(\boldsymbol{1} - p_{switch}) \quad (7)$$

When a novel option was introduced, it could replace the chosen stimulus, or one of the other two stimuli. In this case if the chosen target, *i*, was not rewarded and a different target, *j*, was replaced, we have

$$p_t(\ldots, C_i + \boldsymbol{1}, R_i, C_j = \boldsymbol{0}, R_j = \boldsymbol{0} \mid s_t = [\ldots, C_i, R_i, \ldots], a = choose\ i) = (\boldsymbol{1} - q_i)p_{switch}/\boldsymbol{3} \quad (8)$$

And if the chosen target was not rewarded and was replaced

$$p_t(\ldots, C_i = \boldsymbol{0}, R_i = \boldsymbol{0}, \ldots \mid s_t = [\ldots, C_i, R_i, \ldots], a = choose\ i) = (\boldsymbol{1} - q_i)p_{switch}/\boldsymbol{3} \quad (9)$$

And correspondingly, following a reward and replacement of a different target

$$p_t(\ldots, C_i + \boldsymbol{1}, R_i + \boldsymbol{1}, C_j = \boldsymbol{0}, R_j = \boldsymbol{0} \mid s_t = [\ldots, C_i, R_i, \ldots], a = choose\ i) = q_i p_{switch}/\boldsymbol{3} \quad (10)$$

And

$$p_t(\ldots, C_i = \mathbf{0}, R_i = \mathbf{0}, \ldots \mid s_t = [\ldots, C_i, R_i, \ldots], a = choose\ i) = q_i p_{switch}/\mathbf{3} \quad (11)$$

Note that when a novel option is substituted for the chosen stimulus, the same subsequent state is reached with or without a reward.

Magnetic Resonance Imaging (MRI) Acquisition, Processing, and Analysis.

Image acquisition.—All MRI scans were acquired on a 3T Siemens Tim Trio system with a 32-channel phased-array head coil. T1-weighted (T1w) structural MRI was acquired via a multi-echo MPRAGE sequence (5-echo; voxel size=1mm iso). T2*-weighted functional MRI data were acquired with a gradient EPI pulse sequence using simultaneous multi-slice technology (TR=1s; TE=30ms; Flip=44°; MB factor=4; voxel size=3mm iso). Acquired data were converted from DICOM to Brain Imaging Data Structure (BIDS) format using Heudiconv v.0.5.4 (https://heudiconv.readthedocs.io/en/latest/).

Image preprocessing.—Results included in this manuscript come from preprocessing performed using fMRIPrep 20.2.0rc0 (Esteban et al., 2019) (RRID:SCR_016216), which is based on Nipype 1.5.1 (Gorgolewski et al., 2011) (RRID:SCR_002502).

fMRI analysis.

First-level model.: Participant-level fMRI data were modeled as a function of parameters from the Partially Observable Markov Decision Process (POMDP) model. Specifically, in a first pass model (Model 1) we fit the fMRI timeseries as a function of seven regressors: 1) a choice constant, 2) exploration bonus (BONUS), 3) immediate expected value (IEV), 4) future expected value (FEV), 5) number of trials since a novel stimulus was inserted, 6) whether or not the previous trial was rewarded, and 7) reward prediction error (feedback 1 or 0, minus IEV). Additionally, to isolate BONUS variance components sensitive to the novelty of stimuli in the choice set, we ran an additional model where we did not covary for regressor #5 (i.e., “number of trials since novel”; Model 2; Figure 4). We then ran a group-level comparison computing the *posterior distribution difference* between the BONUS parameter estimates between the original model and the revised model, reasoning that brain regions sensitive to novelty will vary their activations based on whether or not this “number of trials since novel” regressor was included as a covariate at the first-level.

The choice constant was fit across the full 2s duration of the stimulus presentation event. Parametric modulation of the BOLD signal by regressors 2–6 at the time of choice was modeled with duration 0, and we accounted for variation in the BOLD responses using the default FMRIB Linear Optimal Basis Sets (FLOBS) in FSL (Smith et al., 2004). The default FLOBS set comprises three waveforms: a canonical hemodynamic response function (HRF), and its temporal and dispersion derivatives. For each event regressor, the canonical HRF was orthogonalized to the derivative waveforms, and only the HRF parameter estimates were used in second-level models. Regressor 7 was fit the same way, but was time-locked to the feedback event. Notably, multicollinearity was not a concern in either model (all VIFs≤1.54). Cue onset timing was jittered and optimized via AFNI’s make_random_timings.py. 24 standard head motion parameters and their derivatives

were included as confound regressors(Satterthwaite et al., 2013), and 5mm full-width half-maximum smoothing and 100s high-pass temporal filtering were applied to the first-level data.

Second-level model.: Second-level data were modeled using a Bayesian Multilevel Modelling approach using brms v2.15.0 (https://cran.r-project.org/web/packages/brms/index.html) in R v4.0.4 (https://www.r-project.org). Relative to conventional mass univariate analyses, Bayesian Multilevel Modelling improves model efficiency and sensitivity for detecting effects at smaller brain regions (Chen et al., 2019). First, mean percent signal change was extracted for each participant from 370 anatomically-defined regions-of-interest (ROIs) spanning cortex and subcortex. Specifically, cortex was parcellated using 360 areas (180 from each hemisphere) from the Glasser Multimodal Parcellation (Glasser et al., 2016), and 10 subcortical ROIs comprising amygdala and basal ganglia were segmented using the Harvard-Oxford Probabilistic Atlas via the FMRIB Software Library (FSL; (Smith et al., 2001, 2004)). Next, second-level models were fit in brms for each first-level model parameter of interest in the form:

$$Y \mid se \sim 1 + (1 \mid subject) + (1 \mid ROI) \tag{12}$$

In these models, Y corresponds to the percent signal change to a given regressor from the first-level models and se corresponds to the standard error of this response variable across voxels within each ROI. The $subject$ term represents the random effect associated with each subject, and the ROI term represents the random effect associated with each cortical and subcortical ROI in the model. Models used 4 Markov Chain Monte Carlo chains with 10,000 iterations per chain, and the convergence criterion was $\hat{R} < 1.1$ (all $\hat{R}$ values were $\rightarrow 1$). The only model that demonstrated issues with convergence was IEV, and this was resolved by repeating the IEV model with 20,000 iterations per chain (Supplementary Figure 3). All models used weakly informative priors that are defaults in brms—i.e., a Student's t-distribution with scale 3 and 10 degrees of freedom (Bürkner, 2017).

Lastly, we extracted the marginal posteriors associated with each ROI. Importantly, the main output of each brms model is one overall posterior distribution that is a joint distribution across participants and ROIs in a high-dimensional parameter space, and therefore correction for multiple comparisons is not appropriate (Gelman et al., 2012; Limbachia et al., 2021). Statistical inferences regarding the credibility of each ROI encoding a given regressor were made based on the proportion of each distribution that was above 0 (henceforth, $P+$). $P+$ values less than 0.15 were used to indicate credible evidence for negative encoding of a given regressor, whereas $P+$ values above 0.85 indicated credible evidence for positive encoding. Though "strong," "moderate," and "weak" labels are sometimes assigned to arbitrary $P+$ levels within a Bayesian multilevel modelling framework, our goal in the current study was to map the brainwide computational architecture of explore-exploit decision making for the first time in humans. Therefore, even ROIs demonstrating "weak" positive or negative encoding of a given regressor could be theoretically important, and therefore worthy of discussion in the main text (Limbachia et al., 2021).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

The current human subjects work was supported by the National Institute of General Medical Sciences (NIGMS; P30GM122734). The animal work was supported by the NIMH extramural (MH125824 to VDC) and Intramural Research Program of the National Institute of Mental Health (ZIA MH002929). JH's effort while writing this manuscript was supported via NIGMS (P20GM109089). We would like to thank the UNM Center for Advanced Research Computing, supported in part by the National Science Foundation, for providing the research computing resources used in this work. We would also like to thank the organizers of the Computational Cognitive Neuroscience meeting, as well as Twitter, for providing forums that initiated this collaboration.

References

Asaad WF, and Eskandar EN (2008). Achieving behavioral control with millisecond resolution in a high-level programming environment. J. Neurosci. Methods 173, 235–240. [PubMed: 18606188]

Averbeck BB (2015). Theory of choice in bandit, information sampling and foraging tasks. PLoS Comput. Biol 11, e1004164. [PubMed: 25815510]

Averbeck BB, Djamshidian A, O'Sullivan SS, Housden CR, Roiser JP, and Lees AJ (2013). Uncertainty about mapping future actions into rewards may underlie performance on multiple measures of impulsivity in behavioral addiction: evidence from Parkinson's disease. Behav. Neurosci 127, 245–255. [PubMed: 23565936]

Badre D, Doll BB, Long NM, and Frank MJ (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. Neuron 73, 595–607. [PubMed: 22325209]

Bradley MM, and Lang PJ (2020). International Affective Picture System. Encyclopedia of Personality and Individual Differences 2347–2350.

Bürkner P-C (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. J. Stat. Softw 80, 1–28.

Cavanagh JF, Figueroa CM, Cohen MX, and Frank MJ (2012). Frontal theta reflects uncertainty and unexpectedness during exploration and exploitation. Cereb. Cortex 22, 2575–2586. [PubMed: 22120491]

Chen G, Xiao Y, Taylor PA, Rajendra JK, Riggins T, Geng F, Redcay E, and Cox RW (2019). Handling Multiplicity in Neuroimaging Through Bayesian Lenses with Multilevel Modeling. Neuroinformatics 17, 515–545. [PubMed: 30649677]

Choung O-H, Lee SW, and Jeong Y (2017). Exploring Feature Dimensions to Learn a New Policy in an Uninformed Reinforcement Learning Task. Sci. Rep 7, 17676. [PubMed: 29247192]

Cockburn J, Man V, Cunningham W, and O'Doherty JP (2021). Novelty and uncertainty interact to regulate the balance between exploration and exploitation in the human brain. 22.

Cosme D, Flournoy JC, Livingston J, Lieberman MD, Dapretto M, and Pfeifer JH (2021). Testing the adolescent social reorientation model using hierarchical growth curve modeling with parcellated fMRI data.

Costa VD, and Averbeck BB (2020). Primate Orbitofrontal Cortex Codes Information Relevant for Managing Explore-Exploit Tradeoffs. J. Neurosci 40, 2553–2561. [PubMed: 32060169]

Costa VD, Tran VL, Turchi J, and Averbeck BB (2014). Dopamine modulates novelty seeking behavior during decision making. Behav. Neurosci 128, 556–566. [PubMed: 24911320]

Costa VD, Mitz AR, and Averbeck BB (2019). Subcortical Substrates of Explore-Exploit Decisions in Primates. Neuron 103, 533–545.e5. [PubMed: 31196672]

Cox RW (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res 29, 162–173. [PubMed: 8812068]

Daw ND, O'Doherty JP, Dayan P, Seymour B, and Dolan RJ (2006). Cortical substrates for exploratory decisions in humans. Nature 441, 876–879. [PubMed: 16778890]

Djamshidian A, O'Sullivan SS, Wittmann BC, Lees AJ, and Averbeck BB (2011). Novelty seeking behaviour in Parkinson's disease. Neuropsychologia 49, 2483–2488. [PubMed: 21565210]

Domenech P, Rheims S, and Koechlin E (2020). Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex. Science 369.

Ebitz RB, Albarran E, and Moore T (2018). Exploration Disrupts Choice-Predictive Signals and Alters Dynamics in Prefrontal Cortex. Neuron 97, 475.

Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, Kent JD, Goncalves M, DuPre E, Snyder M, et al. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods 16, 111–116. [PubMed: 30532080]

Friedman J, Hastie T, Tibshirani R, and Others (2001). The elements of statistical learning (Springer series in statistics New York).

Gelman A, Hill J, and Yajima M (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. Journal of Research on Educational Effectiveness 5, 189–211.

Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, et al. (2016). A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178. [PubMed: 27437579]

Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, and Ghosh SS (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Front. Neuroinform 5, 13. [PubMed: 21897815]

Hwang J, Mitz AR, and Murray EA (2019). NIMH MonkeyLogic: Behavioral control and data acquisition in MATLAB. Journal of Neuroscience Methods 323, 13–21. [PubMed: 31071345]

Kakade S, and Dayan P (2001). Dopamine bonuses. Adv. Neural Inf. Process. Syst 131–137.

Kidd C, and Hayden BY (2015). The Psychology and Neuroscience of Curiosity. Neuron 88, 449–460. [PubMed: 26539887]

Koller M (2016). robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models. J. Stat. Softw 75, 1–24. [PubMed: 32655332]

Lima Portugal LC, Alves R. de C.S., Junior OF, Sanchez TA, Mocaiber I, Volchan E, Smith Erthal F, David IA, Kim J, Oliveira L, et al. (2020). Interactions between emotion and action in the brain. Neuroimage 214, 116728. [PubMed: 32199954]

Limbachia C, Morrow K, Khibovska A, Meyer C, Padmala S, and Pessoa L (2021). Controllability over stressor decreases responses in key threat-related brain areas. Commun Biol 4, 42. [PubMed: 33402686]

Logothetis NK, Pauls J, Augath M, Trinath T, and Oeltermann A (2001). Neurophysiological investigation of the basis of the fMRI signal. Nature 412, 150–157. [PubMed: 11449264]

Mansouri FA, Freedman DJ, and Buckley MJ (2020). Emergence of abstract rules in the primate brain. Nat. Rev. Neurosci 21, 595–610. [PubMed: 32929262]

Mitz AR (2005). A liquid-delivery device that provides precise reward control for neurophysiological and behavioral experiments. J. Neurosci. Methods 148, 19–25. [PubMed: 16168492]

Mitz AR, Tsujimoto S, Maclarty AJ, and Wise SP (2009). A method for recording single-cell activity in the frontal-pole cortex of macaque monkeys. J. Neurosci. Methods 177, 60–66. [PubMed: 18977387]

Ogasawara T, Sogukpinar F, Zhang K, Feng Y-Y, Pai J, Jezzini A, and Monosov IE (2021). A primate temporal cortex–zona incerta pathway for novelty seeking. Nat. Neurosci 25, 50–60. [PubMed: 34903880]

Poggio T (1990). A Theory of How the Brain Might Work. Cold Spring Harb. Symp. Quant. Biol 55, 899–910. [PubMed: 2132866]

Puterman ML (1994). Markov Decision Processes. Wiley Series in Probability and Statistics.

Satterthwaite TD, Elliott MA, Gerraty RT, Ruparel K, Loughead J, Calkins ME, Eickhoff SB, Hakonarson H, Gur RC, Gur RE, et al. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. Neuroimage 64, 240–256. [PubMed: 22926292]

Shmuel A, Augath M, Oeltermann A, and Logothetis NK (2006). Negative functional MRI response correlates with decreases in neuronal activity in monkey visual area V1. Nature Neuroscience 9, 569–577. [PubMed: 16547508]

Smith S, Bannister PR, Beckmann C, Brady M, Clare S, Flitney D, Hansen P, Jenkinson M, Leibovici D, Ripley B, et al. (2001). FSL: New tools for functional and structural brain image analysis. NeuroImage 13, 249.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 Suppl 1, S208–S219. [PubMed: 15501092]

Tang H, Costa VD, Bartolo R, and Averbeck BB (2022). Differential coding of goals and actions in ventral and dorsal corticostriatal circuits during goal-directed behavior. Cell Rep. 38, 110198. [PubMed: 34986350]

Tsujimoto Satoshi, and Genovesio Aldo. 2017. Firing Variability of Frontal Pole Neurons During a Cued Strategy Task. Journal of Neuroscience 29 (1): 25–36.

Wilson RC, Wang S, Sadeghiyeh H, and Cohen JD (2020). Deep exploration as a unifying account of explore-exploit behavior.

Wittmann BC, Daw ND, Seymour B, and Dolan RJ (2008). Striatal Activity Underlies Novelty-Based Choice in Humans. Neuron 58, 967–973. [PubMed: 18579085]

Yin L, Xu X, Chen G, Mehta ND, Haroon E, Miller AH, Luo Y, Li Z, and Felger JC (2019). Inflammation and decreased functional connectivity in a widely-distributed network in depression: Centralized effects in the ventral medial prefrontal cortex. Brain Behav. Immun 80, 657–666. [PubMed: 31078690]

Yuen KK (1974). The Two-Sample Trimmed t for Unequal Population Variances. Biometrika 61, 165.

Highlights:

- Value and uncertainty direct explore-exploit decisions in humans and monkeys.
- A prefrontal subdivision unique to primates encodes when exploration is valuable.
- Frontoparietal brain regions show dissociable encoding of value and uncertainty.
- Motivational brain regions complement prefrontal contributions to exploration.

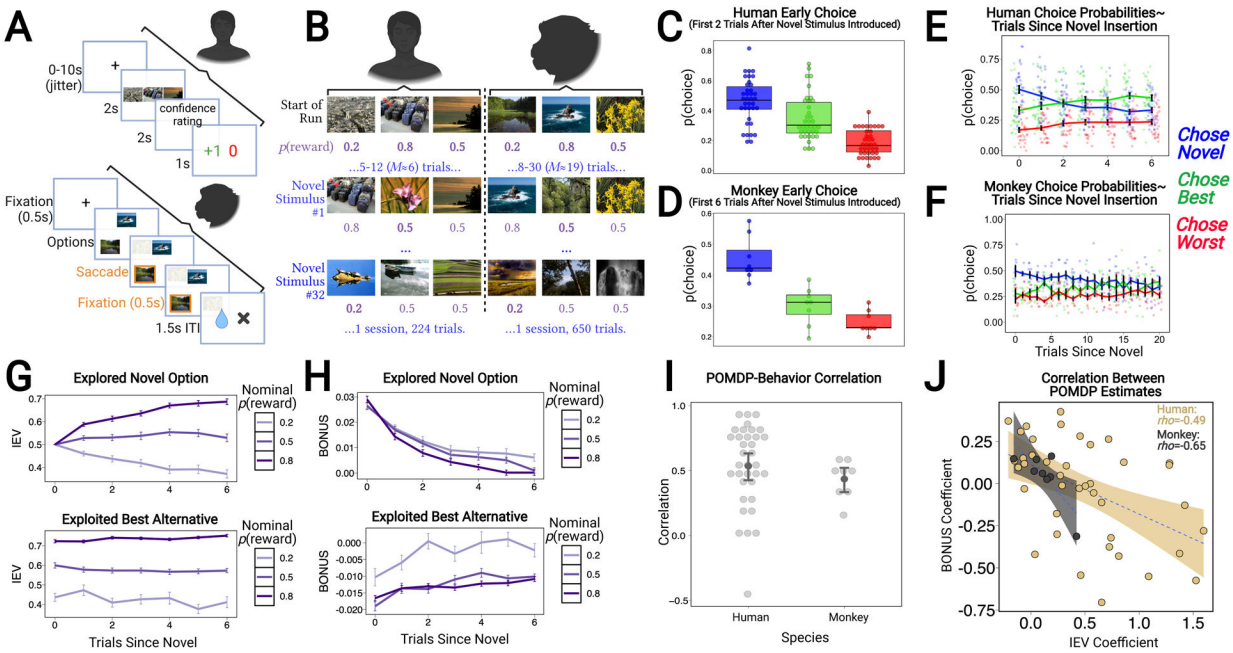

Figure 1. Task Design, Behavioral & Model Performance.

(A) Trial events from the three-arm bandit task, as performed by humans and monkeys (STAR Methods). **(B)** Both humans and macaques chose between neutral images assigned the same nominal reward probabilities, and experienced the same number of overall novel stimulus insertion trials. Insertion rate was faster in humans. **(C-D)** Across both species, the novel stimulus was explored more often than the familiar alternatives were exploited, during the first few trials after a novel stimulus was introduced. When not exploring the novel option, both species exploited the best alternative more often than choosing the worst available option. **(E-F)** Both humans and monkeys selected the novel option less often as the number of trials elapsed since it was introduced and conversely, increased their selection of the best available option. **(G-H)** Mean trial-by-trial changes in the POMDP valuations of human participants' choices broken out by the nominal reward probabilities assigned to each option. The mean IEV and exploration BONUS are shown for when participants explored a novel option (top) versus exploited the best available alternative (bottom). **(I)** The correlation between choice performance and the POMDP was greater than zero within humans and monkeys, and correlation strength did not differ, suggesting similar computations shape explore-exploit behavior between species. (**J**) The parameter estimates used to weight IEV and exploration BONUS were negatively associated across both humans and monkeys.

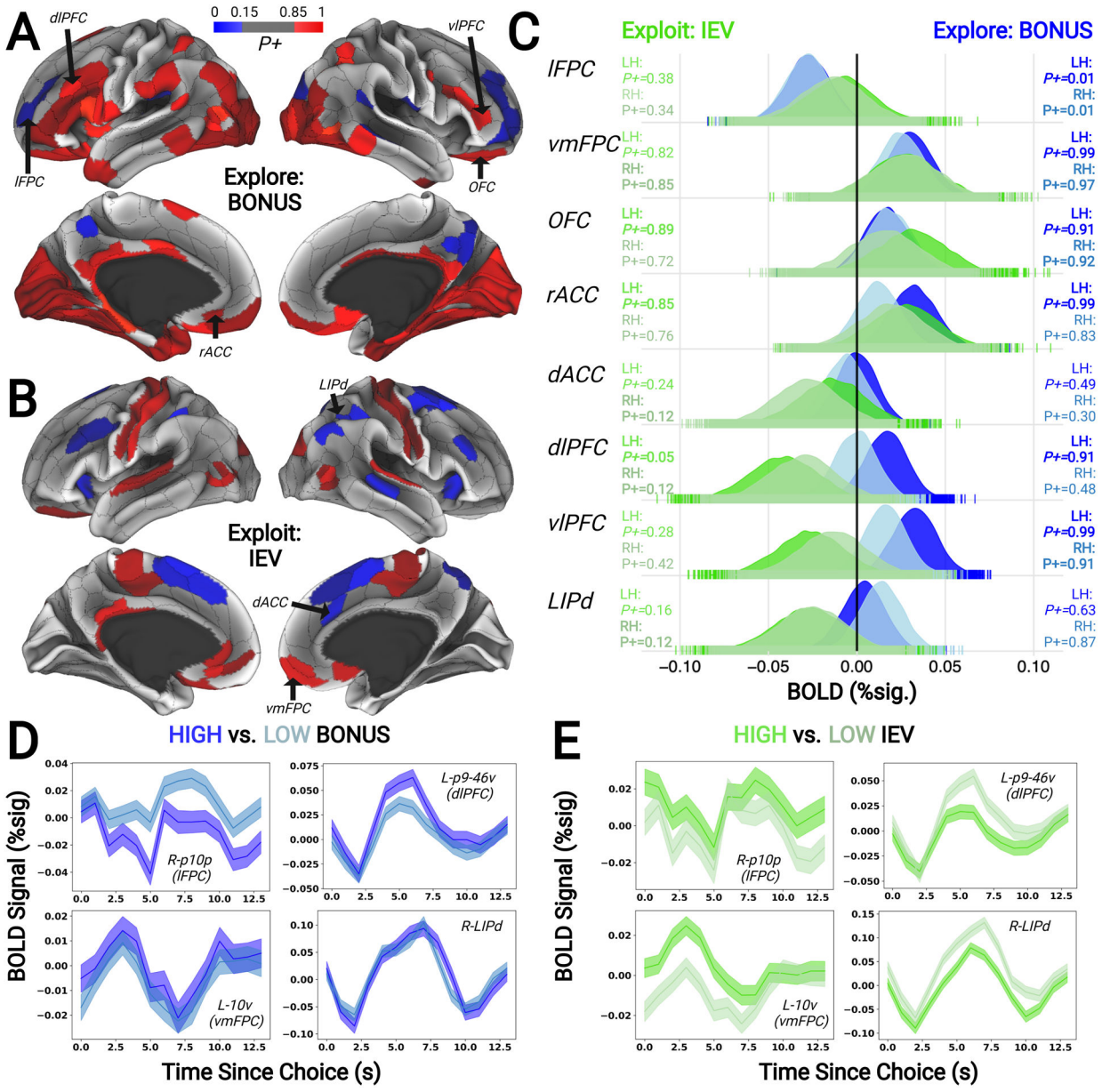

Figure 2. Cortical Encoding of Explore and Exploit Computations.

(A) Cortical regions showing positive (red; ≥85% samples above 0) and negative (blue; ≤15% samples above 0) encoding of a novelty-driven exploration computation (i.e., BONUS) and **(B)** an exploit-related computation (i.e., IEV) in the Bayesian multilevel model. **(C)** Posterior distributions from a subset of *a priori* regions-of-interest indicated that lFPC negatively encoded the BONUS associated with choices, suggesting reduced activation during novelty-driven exploration. Additionally, vmFPC, OFC, and rACC positively encoded both BONUS and IEV parameters suggesting *enhanced* activation during exploration and exploitation. Finally, several frontoparietal network regions demonstrated intraregional dissociations in encoding of PODMP derived value estimates: positive encoding of BONUS and negative encoding of IEV. Darker and lighter colors in posterior distributions indicate left and right hemispheres. Bolded text indicates either ≥85% or ≤15%

of posterior samples above 0. *X*-axis corresponds to BOLD percent signal change; *Y*-axis corresponds to the posterior densities from the Bayesian multilevel model. Time courses from representative ROIs in the Bayesian model showing mean activation when trials are based on splitting the BONUS (**D**) and IEV (**E**) around breakpoints (i.e., 0 for BONUS and 0.5 for IEV). Although, the primary Bayesian multilevel models included parameter estimates that reflected signal modulation as a continuous function of BONUS and IEV. Exploration-related deactivation was observed in lFPC, alongside weak encoding of the IEV of choices (upper left panels in **D** and **E**), while increases in the value of exploring or exploiting increased activation of vmFPC (lower left panels in **D** and **E**). Abbreviations: lateral frontopolar cortex, lFPC; ventromedial frontopolar cortex, vmFPC; orbitofrontal cortex, OFC; rostral anterior cingulate cortex, rACC; dorsal anterior cingulate cortex, dACC; dorsolateral prefrontal cortex, dlPFC; ventrolateral prefrontal cortex, vlPFC; and dorsal lateral intraparietal area, LIPd.

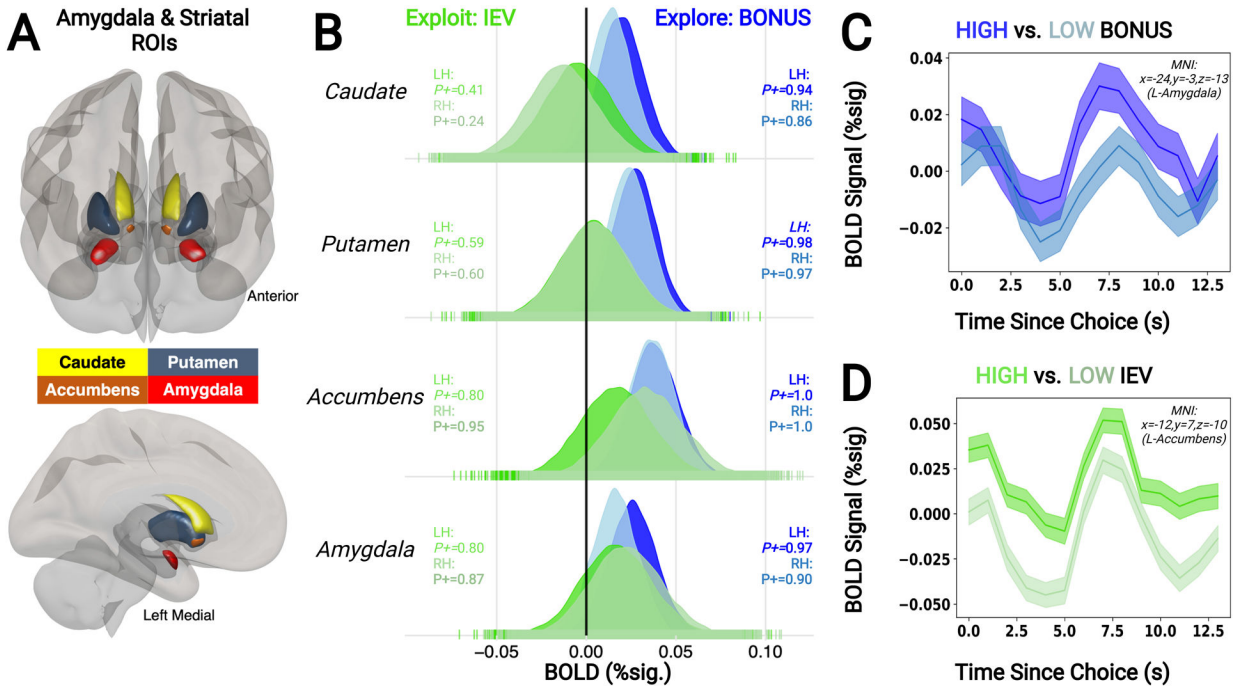

Figure 3. Amygdala & Striatum Encode Explore-Exploit Related Computations.

(A) Amygdala and striatal ROIs included in the Bayesian multilevel model. **(B)** Dorsal striatal nuclei (caudate and putamen) demonstrated evidence for positive encoding of BONUS but not IEV, while ventral nuclei—namely, accumbens and amygdala—demonstrated positive encoding of *both* BONUS and IEV. Darker and lighter colors in posterior distributions indicate left- and righth-emispheres, respectively. Bolded text indicates either ≥85% or ≤15% of posterior samples above 0. **(C-D)** Timecourses pulled from representative clusters within amygdala **(C)** and accumbens **(D)** that demonstrated strong evidence for BONUS and IEV encoding, respectively.

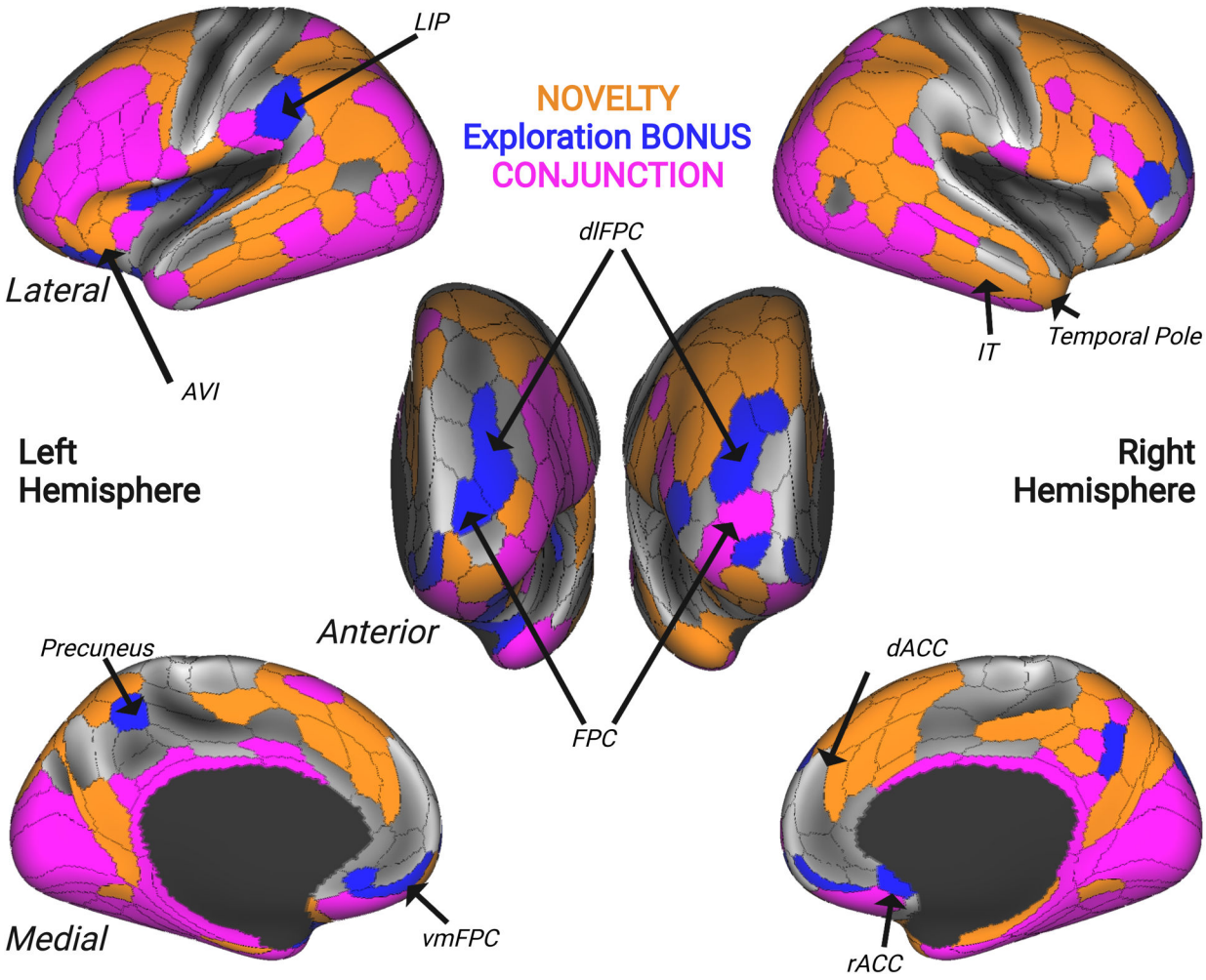

Figure 4. Overlapping and Distinct Networks Represent Novelty and Exploration.

Surface map displaying areas that uniquely encoded the number of trials since a novel stimulus was presented (orange), the relative future value of the selected option (BONUS; blue), and regions that encoded both regressors (CONJUNCTION; pink). Novelty-related encoding was observed in areas of the temporal pole (right TGd), inferior temporal cortex (IT), and regions of the salience network comprising anteroventral insula (AVI) and dorsal anterior cingulate cortex (dACC). BONUS-related encoding was observed in rostral frontopolar cortex (FPC), dorsolateral and ventromedial frontopolar regions (dlFPC and vmFPC), rostral anterior cingulate cortex (rACC), precuneus, and lateral intraparietal area (LIP). Lastly, the conjunction revealed a diffuse network of brain regions that were sensitive to both BONUS value and stimulus novelty, comprising several lateral prefrontal and frontopolar regions, orbitofrontal cortex, mid-insula, and several inferior temporal cortex regions.

Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: Organisms/strains		
Rhesus macaques (Macaca mulatta)	NIMH/NIH	
Original Code		
All original code	This paper	10.5281/zenodo.6342173
Software and algorithms		
AFNI	(Cox, 1996)	https://afni.nimh.nih.gov/
Connectome Workbench	Human Connectome Project	https://www.humanconnectome.org/software/get-connectome-workbench
E-Prime	Psychology Software Tools	https://pstnet.com/products/e-prime/
FMRIB Software Library (FSL)	(Smith et al., 2004)	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki
MATLAB	Mathworks	https://www.mathworks.com/products/matlab.html
Monkeylogic	(Asaad and Eskandar, 2008)	https://www.brown.edu/Research/monkeylogic/
Python	Python Software Foundation	https://www.python.org/download/releases/3.0/
R	The R Foundation	https://www.r-proiect.org/