Running head: NEURAL CORRELATES OF EXPLORE-EXPLOIT

Title: The neurocomputational architecture of explore-exploit decision making.

Authors: Jeremy Hogeveen[1*], Teagan S. Mullins[1], John Romero[1], Elizabeth Eversole[1], Kimberly Rogge-Obando[1], Andrew R. Mayer[1-4], & Vincent D. Costa[5*]

Affiliations:

1. Department of Psychology & Psychology Clinical Neuroscience Center,
2. Department of Psychiatry & Behavioral Sciences,
3. Department of Neurology, University of New Mexico, Albuquerque, NM, USA.
4. Mind Research Network/LBERI, Albuquerque, NM, USA.
5. Department of Behavioral Neuroscience, Oregon Health and Science University, Portland, OR, USA.

*: Co-corresponding authors (jhogeveen@unm.edu; costav@ohsu.edu)

Keywords: Reward, explore-exploit, novelty-seeking, fMRI, computational modelling, frontopolar cortex.

Funding: The current human subjects work was supported by the National Institute of General Medical Sciences (NIGMS; P30GM122734), and the animal work was supported by the Intramural Research Program of the National Institute of Mental Health (NIMH; ZIA MH002929). JH's effort while writing this manuscript was supported via NIGMS (P20GM109089).

Abstract

People often make the difficult decision to try new options (exploration) and forego immediate rewards (exploitation). Novelty-seeking is an adaptive solution to this explore-exploit dilemma that has been studied using targeted recordings in monkeys, but our understanding of the neural computations supporting novelty-seeking in humans is limited. Here, we show homologous computations supporting novelty-seeking across humans and monkeys, and reveal a previously unidentified cortico-subcortical architecture mediating explore-exploit behavior in humans.

Body

The motivation to explore novel information instead of resorting to already learned behaviors is a bedrock for how humans learn across the lifespan. Yet, 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, and these decisions are implemented via interactions between prefrontal and motivational brain regions (e.g. amygdala and striatum). However, considerable controversy remains regarding the neural computations that drive exploration. In one view, prefrontal cortex and motivational regions may work together to compute both the anticipated immediate and future value of choice options, driving the exploration of novel opportunities when it could lead to more rewards in the future[1–3]. In another view, prefrontal cortex may override encoding of familiar outcomes in motivational regions while forming new decision policies[4–7]. Resolving the neurocomputational architecture of explore-exploit decisions could contribute to the development of more effective circuit-based treatments for transdiagnostic psychiatric challenges including inflexibility[8]. Here, we probe the computations encoded in prefrontal and motivational regions during explore-exploit behavior in humans using model-based fMRI, and build a translational bridge for future invasive and noninvasive studies by demonstrating homologous explore-exploit computations across humans and monkeys on the same task.

The neurocomputational bases of explore-exploit decisions in primates have been investigated using neural recordings in monkeys. A network comprising amygdala, ventral striatum, and orbitofrontal cortex encodes *both* the anticipated immediate *and* latent future value of choice options during explore-exploit decisions[1,3]. However, these studies are restricted to *a priori* targeted recording sites and therefore limited in anatomical scope. Whole brain examinations of explore-exploit decisions via fMRI in humans have suggested that lateral frontopolar cortex (lFPC) and posterior parietal regions also play a role in explore-exploit decisions[4,9,10]. However, in these studies 'exploration' is typically defined as an overriding of 'exploitative' decision policies (i.e., failure to maximize value), making it difficult to determine whether FPC is encoding computations related to goal-directed exploration, random exploration as a function of decision noise, or poor reinforcement learning[11]. In the current study, human and monkey subjects performed a reinforcement learning task involving periodic insertion of novel stimuli with an unknown value. Novelty-seeking in this context is an evolved solution to the explore-exploit dilemma, which should explicitly signal a motivation to engage in goal-directed exploration. Importantly, recent advances in computational modelling with a normative agent framework enabled us to formally quantify the immediate and latent future value for each option, independent of observed behavior[1,3,12]. To be clear, this allowed us to directly compare the value of exploring versus exploiting within the same brain regions—and across the whole brain.

We utilized a three-arm bandit task where novelty was used to motivate exploration (**Figure 1A**). Specifically, monkeys and humans performed speeded decisions between 3 neutral images assigned low (p_{reward}=0.2), medium (p_{reward}=0.5), or high (p_{reward}=0.8) reward values. Periodically, a novel stimulus with a randomly assigned value was inserted, forcing participants to face a tradeoff between exploring the new option versus exploiting the best available alternative. Both monkeys and human participants were more likely to explore the novel stimulus on the early[1] trials post-insertion (Monkey: *M*=0.45, *SEM*=0.02; Human: *M*=0.46, *SEM*=0.02) than to exploit the best alternative option (Monkey: *M*=0.31, *SEM*=0.02, t_{yuen}=3.27,

[1]"Early trials" were defined as the first *N*=2 trials post-insertion for humans. Relative to the typical interval between insertion trials (*M*≈6 for humans, *M*≈19 for monkeys), this was translated to *N*=6 post-insertion trials for the monkeys.

p=0.02, η=0.88; Human: *M*=0.35, *SEM*=0.02, t_{yuen}=2.96, *p*=0.007, η=0.55; **Figure 1B-C**). In turn, the tendency to exploit the best alternative option on post-insertion trials was higher than selection of the worst alternative (Monkey: *M*=0.24, *SEM*=0.01, t_{yuen}=3.48, *p*=0.02, η=0.70; Human: *M*=0.19, SEM_{human}=0.01, t_{yuen}=3.93, *p*<0.001, η=0.79; **Figure 1B-C**). Looking at decision making over time, the probability of selecting the novel stimulus decreased as the number of trials post-insertion 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 1D-E**), 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 1D-E**). Collectively, behavioral performance indicated two homologous patterns across primate species: i) a novelty-seeking bias when making explore-exploit decisions, and ii) in time, this novelty bias wanes and primates learn to exploit the option with the highest assigned value.

We modeled normative explore-exploit policies as a Partially Observable Markov Decision Process (POMDP). In the POMDP, each option's value is the sum of its immediate expected value (IEV) and future expected value (FEV). IEV reflects the likelihood that a particular choice will be rewarded. FEV is latent and reflects the sum of potential rewards to be earned in the future. Trial-to-trial changes in novelty-seeking are thought to be primarily shaped by an exploration BONUS, the relative difference in the FEV of an individual option relative to the FEV of all options. The exploration BONUS is highest when a novel option is first introduced, and decreases each time an option is sampled.

The POMDP value estimates were used in a non-linear regression approach to predict which option humans or monkeys would choose. In combination, the IEV and exploration BONUS associated with choices made by humans (*M*=0.59, *95%CI*=0.47 to 0.70, *t*=10.5, *p*<0.001) or rhesus macaques (*M*=0.45, *95%CI*=0.30 to 0.61, *t*=7.43, *p*<0.001) were predictive of overall performance. Examination of the individual regression coefficients in each species indicated that the IEV associated with each option was a strong determinant of whether or not an option was chosen (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). In humans, the exploration BONUS was not as strong as a predictor as it was 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), perhaps because novelty was more salient to the monkeys due to differences in the number of trials that elapsed between the introduction of novel choice opportunities. However, in both humans and monkeys, we observed a negative correlation between IEV and BONUS regression coefficients (Human: *rho*=-0.48, *95%HDI*=-0.73 to -0.19; Monkey: *rho*=-0.67, *95%HDI*=-0.97 to -0.04). The strength of the POMDP-Behavior correlation did not differ across primate species (M_{diff}=0.13, *95%CI*=-0.03 to 0.30, *t*=1.64, *p*=0.09; **Figure 1F**).

To elucidate the neurocomputational architecture of explore-exploit decision making in humans, we collected fMRI data and modeled choice-evoked responses as a function of value estimates, IEV and BONUS, derived from the POMDP model[3,12]. Bayesian multi-level modeling[13] was used to identify cortical and subcortical regions that encoded trial-by-trial changes in exploration BONUS and IEV associated with participants' choices —given their importance in shaping exploration[14] and exploitation[15], respectively. Several cortical and subcortical regions-of-interest (ROIs) encoded both relative IEV and BONUS, suggesting a role for these regions in shaping both exploitation and novelty-driven exploration. Both relative IEV and BONUS were associated with enhanced activation of medial FPC (mFPC), lateral OFC, subgenual anterior cingulate cortex (sgACC; **Figure 2**), and ventral subcortical regions (accumbens and amygdala; **Figure 3**). Conversely, both relative IEV and BONUS were associated with reduced activation of lateral FPC (lFPC; **Figure 2**).

In contrast, several regions demonstrated dissociable encoding of exploit- and explore-related computations. Specifically, whereas frontoparietal network regions (namely, dorsolateral prefrontal cortex, dlPFC; ventrolateral prefrontal cortex, vlPFC; area lateral intraparietal ventral, LIPv; and dorsal anterior cingulate cortex, dACC) demonstrated positive encoding of exploration BONUS, they negatively scaled with relative IEV (**Figure 2**). Dorsal striatal nuclei (caudate and putamen) positively encoded BONUS, and caudate negatively scaled with relative IEV (**Figure 3**). Sensorimotor responses also diverged across exploit- and explore-related computations: Visual areas showed enhanced activation as a function of BONUS and reduced activation as a function of relative IEV, whereas primary somatomotor cortices demonstrated enhanced activation as a function of relative IEV. For a complete description of the fMRI analysis pipeline and results, see **Supplementary Materials**.

The current results provide key insight into the neural computations underlying explore-exploit decision making in primates. Homologous to recent monkey neurophysiological data[3,16], we found encoding of explore- and exploit-related computations in regions known to be essential for reward-guided decision making—namely, ventral striatum[cf.14], amygdala, and ventromedial prefrontal regions. This suggests that homologous cortico-subcortical motivational neural circuits help to drive both reward-guided and novelty-seeking behaviors across primate species. Importantly, evidence for homologous neural computations shaping novelty-driven exploration across monkeys and humans should inform the development of next generation treatments for mental disorders associated with pathological explore-exploit behavior.

POMDP derived valuations related to explore or exploit were negatively encoded in lFPC. These effects were driven by greater deactivation of lFPC when participants exploited choice options they had repeatedly sampled and that they had learned were rewarded at a relatively high rate, *and* when participants explored novel options with an unknown probability of reward. Conversely, this implies increased activation of lFPC when participants chose low value options that had not yet been sampled or sampled infrequently several trials after a novel stimulus was last inserted (i.e., choices with low BONUS *and* low-to-moderate relative IEV), likely indicating increased recruitment of lFPC during late-run exploration in search of better choice opportunities. In human fMRI studies, lFPC is known to be activated in tasks that require multi-step planning or the control of responses according to competing goals[17,18]. Negative encoding of BONUS and relative IEV at the time of choice in lFPC may indicate that this region plays a role in outcome monitoring and retaining the value of *unchosen* options[10]. An lFPC mechanism for outcome monitoring and encoding the value of unselected options at the time of choice could be critical for enabling primates' remarkable ability to defer goals to explore new alternatives[19].

Lastly, we observed dissociations between explore-exploit computations in dorsal corticostriatal brain networks. Specifically, in the frontoparietal network and dorsal striatal nuclei (caudate and putamen), BOLD activity was increased as a function of BONUS, suggesting a role in computing the latent value of exploring novel stimuli. Several of these frontoparietal and dorsal striatal regions were also *negatively* associated with relative 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, and the presence of salient periodic novel stimulus insertions. 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 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 known alternatives post-insertion trials (when all options are familiar, and their

relative IEV is differentiated), participants may begin making decisions using a model-free stimulus-outcome learning system. Therefore, positive encoding of BONUS and negative encoding of relative IEV in dorsal frontoparietal and dorsal striatal regions would accord with the view that these regions play a role in Bayesian state inference and model-based explore-exploit decision making[20]. Collectively, these findings should motivate future single-cell recording studies in lFPC, frontoparietal circuits, and dorsal striatum during explore-exploit behavior in nonhuman primates.

Overall, the current study provides compelling evidence that humans and monkeys perform similar neural computations when exploring novel stimuli in lieu of exploiting familiar rewards. Specifically, across primate species we observed a novelty-seeking decision bias in the immediate wake of a novel stimulus presentation, alongside an increased tendency to exploit the best available option relative to the worst available alternative. These decision tendencies were well-modeled as a POMDP across both human participants and monkey subjects, with the model fit not differing significantly across species. Therefore, novelty-seeking under explore-exploit tensions 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 (e.g. substance use disorder).

Figures

Figure 1. **(A)** Details of the Novelty Bandit Task session as performed by human and monkey. Both animals 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. **(B-C)** Across both species, selection of the novel stimulus was increased relative to selection of the best available alternative on the early trials post-novel stimulus insertion. Both species also exploited the best alternative more often than choosing the worst available option on early trials post-insertion. **(D-E)** Both humans and monkeys decreased their sampling of the novel option over time as a function of trials post-novel stimulus insertion, and conversely both species also increased their selection of the best available option over time. **(F)** The correlation between behavioral task performance and the POMDP was significantly greater than zero within humans and monkeys, and POMDP-Behavior correlation strength did not differ between humans and monkeys, suggesting similar computations shape explore-exploit behavior across primate species.

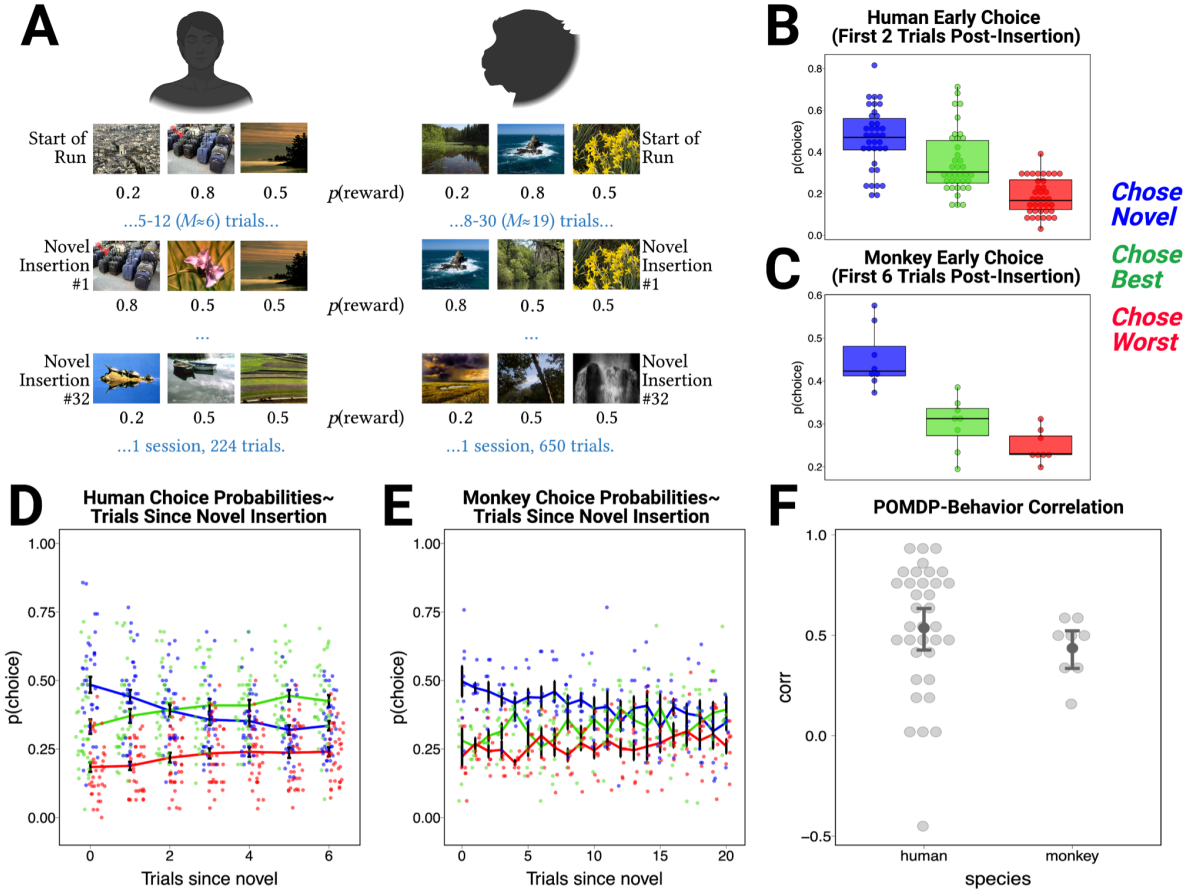


Figure 2. **(A)** Cortical regions showing credible evidence for positive (Red; ≥85% samples above 0) and negative (Blue; ≤15% samples above 0) encoding of an exploit-related computation (i.e., relative immediate expected value, IEV) **(B)** and an exploration / novelty-seeking computation (i.e., BONUS) in the Bayesian multilevel model. **(C)** Posterior distributions from subset of *a priori* regions-of-interest indicated a dissociation in frontopolar cortex between lateral FPC, which negatively encoded both relative IEV and BONUS suggesting reduced activation during both exploration and exploitation, and mFPC, OFC, and sgACC which positively encoded both parameters suggested *enhanced* activation during exploration and exploitation. Frontoparietal regions (namely: dACC, dlPFC, vlPFC, and LIPv) demonstrated within-region dissociations across exploit- and explore-related computations: Positive encoding of BONUS and negative encoding of relative IEV. Darker and lighter colors in posterior distributions indicate left- and right-hemispheres, respectively. Bolded text indicates either ≥85% or ≤15% of posterior samples above 0.

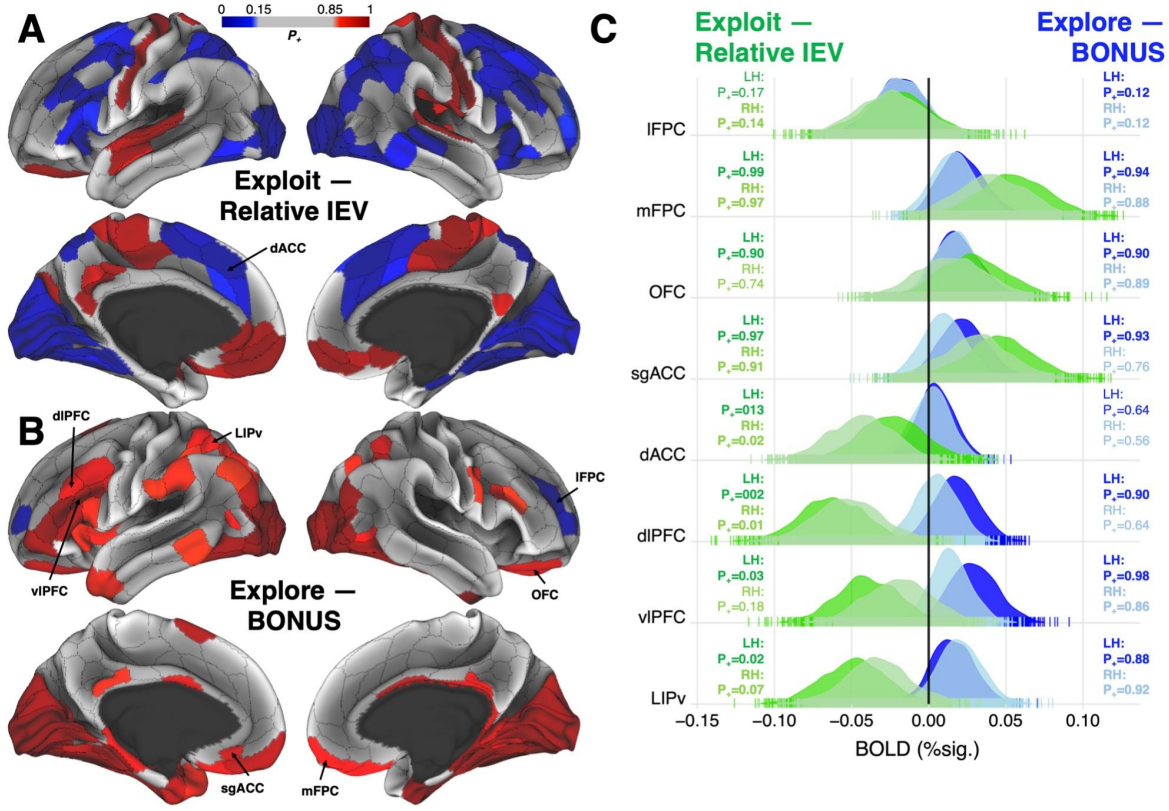
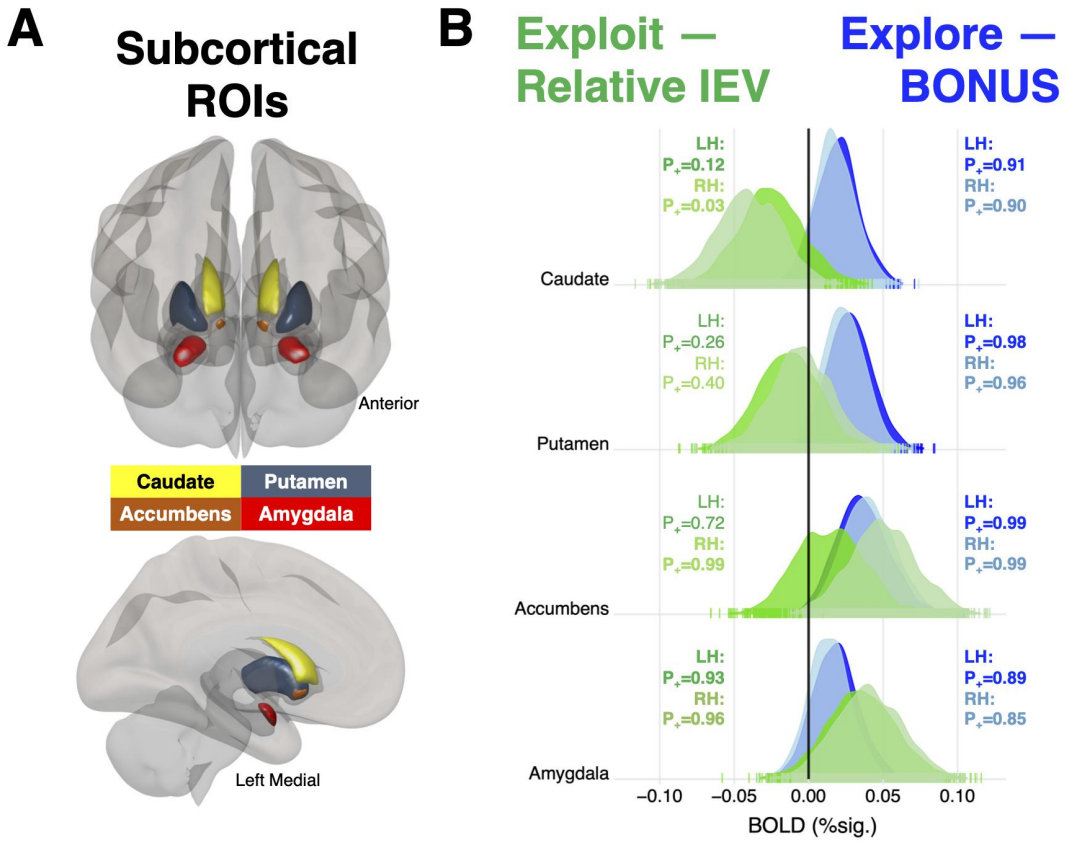


Figure 3. (A) Subset of the subcortical ROIs included in the Bayesian multilevel model. **(B)** Credible (≥85% samples above 0) evidence for positive encoding of relative IEV in accumbens and amygdala, while caudate negatively encoded relative IEV (≤15% samples above 0). In contrast, BONUS was positively encoded in dorsal striatum, accumbens, and amygdala. Darker and lighter colors in posterior distributions indicate left- and right-hemispheres, respectively. Bolded text indicates either ≥85% or ≤15% of posterior samples above 0.



References

1. Costa, V. D. & Averbeck, B. B. Primate Orbitofrontal Cortex Codes Information Relevant for Managing Explore–Exploit Tradeoffs. *The Journal of Neuroscience* vol. 40 2553–2561 (2020).
2. Wilson, R. C., Wang, S., Sadeghiyeh, H. & Cohen, J. D. Deep exploration as a unifying account of explore-exploit behavior. doi:10.31234/osf.io/uj85c.
3. Costa, V. D., Mitz, A. R. & Averbeck, B. B. Subcortical Substrates of Explore-Exploit Decisions in Primates. *Neuron* **103**, 533–545.e5 (2019).
4. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
5. Domenech, P., Rheims, S. & Koechlin, E. Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex. *Science* **369**, (2020).
6. Choung, O.-H., Lee, S. W. & Jeong, Y. Exploring Feature Dimensions to Learn a New Policy in an Uninformed Reinforcement Learning Task. *Scientific Reports* vol. 7 (2017).
7. Ebitz, R. B., Albarran, E. & Moore, T. Exploration Disrupts Choice-Predictive Signals and Alters Dynamics in Prefrontal Cortex. *Neuron* **97**, 475 (2018).
8. Voon, V., Reiter, A., Sebold, M. & Groman, S. Model-Based Control in Dimensional Psychiatry. *Biol. Psychiatry* **82**, 391–400 (2017).
9. Badre, D., Doll, B. B., Long, N. M. & Frank, M. J. Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron* **73**, 595–607 (2012).
10. Boorman, E. D., Behrens, T. E. J., Woolrich, M. W. & Rushworth, M. F. S. 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 (2009).
11. Zajkowski, W. K., Kossut, M. & Wilson, R. C. A causal role for right frontopolar cortex in directed, but not random, exploration. *Elife* **6**, (2017).

12. Averbeck, B. B. Theory of choice in bandit, information sampling and foraging tasks. *PLoS Comput. Biol.* **11**, e1004164 (2015).
13. Chen, G. *et al.* Handling Multiplicity in Neuroimaging Through Bayesian Lenses with Multilevel Modeling. *Neuroinformatics* **17**, 515–545 (2019).
14. Wittmann, B. C., Daw, N. D., Seymour, B. & Dolan, R. J. Striatal Activity Underlies Novelty-Based Choice in Humans. *Neuron* vol. 58 967–973 (2008).
15. Hunt, L. T. *et al.* Mechanisms underlying cortical activity during value-guided choice. *Nature Neuroscience* vol. 15 470–476 (2012).
16. Costa, V. D. & Averbeck, B. B. Primate Orbitofrontal Cortex Codes Information Relevant for Managing Explore-Exploit Tradeoffs. *J. Neurosci.* **40**, 2553–2561 (2020).
17. Koechlin, E., Basso, G., Pietrini, P., Panzer, S. & Grafman, J. The role of the anterior prefrontal cortex in human cognition. *Nature* **399**, 148–151 (1999).
18. Braver, T. S., Reynolds, J. R. & Donaldson, D. I. Neural mechanisms of transient and sustained cognitive control during task switching. *Neuron* **39**, 713–726 (2003).
19. Tsujimoto, S., Genovesio, A. & Wise, S. P. Frontal pole cortex: encoding ends at the end of the endbrain. *Trends in Cognitive Sciences* vol. 15 169–176 (2011).
20. Bartolo, R. & Averbeck, B. B. Prefrontal Cortex Predicts State Switches during Reversal Learning. *Neuron* **106**, 1044–1054.e4 (2020).