LINKING NEURAL REPRESENTATIONS TO ADAPTIVE BEHAVIOR WITH COGNITIVE MODELING

Christina Maher[1], **Salman Qasim**[2], **Lizbeth Nuñez Martinez**[1], **Angela Radulescu**[1*], **Ignacio Saez**[1*]
[1]Icahn School of Medicine at Mount Sinai [2]Rutgers University
christina.maher@icahn.mssm.edu
*Authors contributed equally.

ABSTRACT

Humans efficiently navigate complex learning and decision-making by forming representations of task-relevant information, a process facilitated by selective attention. Although the lateral prefrontal cortex (LPFC) and orbitofrontal cortex (OFC) are linked to attention and value-based decision-making, the neurophysiological processes that coordinate these regions in the maintenance of task-relevant representations remain unclear. To investigate this, we combined intracranial electrophysiology (iEEG) from OFC and LPFC of neurosurgical epilepsy patients with cognitive modeling of behavior, providing the spatiotemporal resolution to test local and circuit-level hypotheses about neural representations. Our findings reveal how shared computational strategies across brain regions and individuals enable the brain to maintain representations critical for adaptive decision-making. This approach offers a novel framework for measuring representational alignment at both neural and subject levels, uncovering the neurocomputational principles that drive real-world behavior. By integrating iEEG and cognitive modeling, we present an approach for studying representational alignment, revealing how it emerges both across brain regions, reflected in shared spectral and temporal features of neural state representations, and across individuals who adopt similar computational strategies to solve real-world decision-making tasks.

1 INTRODUCTION

Every day, we make decisions by evaluating multidimensional options. Learning the value of each option while simultaneously attending to relevant dimensions is a complex task. Yet, humans navigate this complexity with remarkable efficiency, by maintaining task representations that selectively focus on relevant information (Niv, 2019). This ability makes real-world, multidimensional learning and decision-making tractable.

Although neuroimaging has revealed correlates of attention and value learning in the lateral prefrontal cortex (LPFC) (Leong et al., 2017; Miller & Buschman, 2013) and orbitofrontal cortex (OFC) (Saez et al., 2018), respectively, the precise neurophysiological processes underlying state representations in humans remain unclear. To address this, we combined human intracranial electrophysiology (iEEG) recordings from the OFC and LPFC of neursurgical epilepsy patients with cognitive modeling of behavior. iEEG provides the spatiotemporal resolution needed to test local and circuit-level

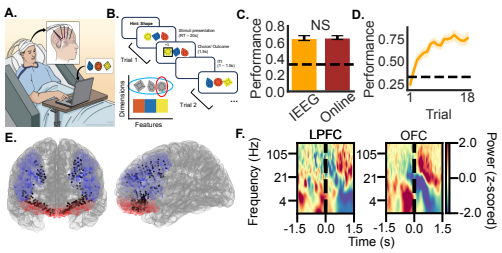

Figure 1: Methods. A. Neurosurgical participants (N=21) performed a B. multidimensional RL task. C. iEEG (N=21) and online (N=50) participants' performance was not significantly different ($t(48) = 0.10$, $p > 0.05$). D. Proportion of correct choices increased across trials (error bar = SEM). E. Electrodes in OFC (red; 159 electrodes) and LPFC (blue; 111 electrodes) across participants. F. LFP signal decomposed by frequency band and normalized using ITI for one participant (2 LPFC, 6 OFC electrodes), time-locked to choice/reward outcome ($t = 0$).

hypotheses about the physiological characteristics of neural representations, while cognitive modeling enables a direct characterization of the brain's otherwise latent decision-making mechanisms. Our results demonstrate that shared computational strategies are manifested in aligned neural representations across biological systems, both at the level of brain regions and individuals (Sucholutsky et al., 2024).

2 METHODS

This novel methods integration offers a powerful framework for measuring representational alignment at both the neural and subject levels, allowing us to uncover the neurocomputational principles that govern the maintenance of state representations essential for adaptive, real-world behavior. By linking cognitive model parameters to spectral and temporal features of neural activity, we identify shared representational formats across brain regions and among individuals who adopt similar learning strategies. Distinct computational strategies for multidimensional RL are reflected in participants' neural representations of the task. We observed representational alignment across individuals, reflected in their model-inferred cognitive strategies and the spectral and temporal characteristics of the circuit-level neural activity that underlie their observed behavior. We also found alignment between discrete brain regions, measured by the similarity in spectral and temporal encoding of model-derived state information.

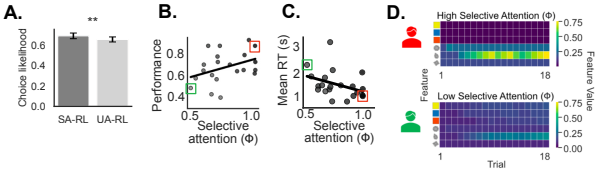

Figure 2: RL modeling results. A. SA-RL explained behavior best ($t(20) = 2.89$, $p < 0.01$; error bars = SEM). B. Selective attention improved task performance ($r(19) = 0.39$, $p < 0.05$) and C. reaction time ($r(19) = -0.44$, $p < 0.05$) by D. modulating value learning. Feature values were computed using the SA-RL model with participants' fitted parameters. Red figure is a participant with higher selective attention (red square, panels B/C): value assignment concentrated on relevant dimension (shape). Green figure is a participant with lower attention (green square, panels B/C): value assignment spread across dimensions.

Multidimensional RL Task: We adapted a multidimensional reinforcement learning task from prior work (Leong et al., 2017; Niv et al., 2015; Wilson & Niv, 2012) for neurosurgical iEEG patients. Participants (N=21) completed six 18-trial games, choosing between three stimuli varying in shape (square, oval, circle) and color (orange, yellow, blue) (Fig. 1A/B). One dimension (shape or color) was relevant per game, and selecting the target feature yielded reward with 80% probability. The relevant dimension was cued between games, and participants were informed of the task structure. iEEG participants performed above chance and comparably to an online control group (Fig. 1C/D).

RL Models: We evaluated two RL models: Uniform Attention RL (UA-RL) and Selective Attention RL (SA-RL). Both models are based on the Rescorla-Wagner learning rule and have been validated in previous work (Maher et al., 2024; Leong et al., 2017). The UA-RL model implements uniform attention to both dimensions of each stimulus, whereas the SA-RL model implements selective attention to the instructed relevant dimension. We assume participants choose based on expected value (EV), computed as $V_t(S_j) = \sum_d \Phi \cdot v_t(d, S_i)$, where Φ is the attention weight on dimension d and $v_t(d, S_i)$ is the value of that feature. After feedback, the reward prediction error is $\delta_t = r_t - V_t(S_c)$, which updates chosen feature values via $v_{t+1}(d, S_c) = v_t(d, S_c) + \eta \cdot \Phi \cdot \delta_t$.

Choice probability was computed using a softmax rule. In the SA-RL model, Φ was a free parameter that biased choice and learning toward the instructed relevant dimension, capturing individual differences in attentional strategy (Fig. 2D). In contrast, the UA-RL model fixed Φ=0.50, assuming uniform attention. Model fit was assessed using leave-one-game-out cross-validation for maximum likelihood estimation.

Intracranial Electrophysiology: We recorded local field potentials from OFC and LPFC (Fig. 1E/F) to examine region-specific neural responses to state features defined by cognitive modeling. High gamma activity (60–200 Hz; HGA), which reflects population-level spatiotemporal dynamics and correlates with single-unit spiking (Rich & Wallis, 2017; Nir et al., 2007), was used to mea-

sure local information encoding. We assessed funtional connectivity using theta activity (2–8 Hz), linked to cross-regional communication and attention (Landau et al., 2015; Fries, 2023). We applied non-parametric multivariate linear models to identify local HGA encoding of behaviorally-relevant features and measured directed cross-regional theta connectivity using Phase Slope Index (PSI).

3 Results and Discussion

Selective Attention Shapes State Representations: To investigate whether participants deploy selective attention during multidimensional RL, we determined UA-RL and SA-RL's ability to explain participants' choices. As hypothesized, the SA-RL model provided a significantly better fit to participants' behavior than the UA-RL model ($t(20) = 2.89, p < 0.01$; Fig. 2A), indicating that state representations during RL are modulated by attention. Moreover, the SA-RL model's attention weight parameter (Φ) captured meaningful behavioral differences (Fig. 2B/C), reflecting individual differences in how participants represent the task environment (Fig. 2D).

Model-based state features encoded in shared neural representations: We examined local encoding of two key RL features: model-free reward and model-based choice EV. To assess the consistency, or representational similarity, of encoding across participants, we ran multivariate regression on each electrode, predicting trial-averaged HGA power from reward outcome, relevant dimension, and chosen features. We generated a null distribution (1000 permutations) and computed z-scored reward β coefficients for each region. Comparing these to an intercept-only model with participant as a random effect, we found significant reward encoding in LPFC HGA, but not in OFC HGA (LPFC: $\beta = -0.70$, $z = -2.10$, $p < 0.05$; OFC: $\beta = -0.19$, $z = -1.11$, $p > 0.05$; Fig. 3A/B). These results suggest that reward encoding in LPFC is relatively homogeneous across participants, giving rise to a consistent group-level effect. In contrast, OFC does not exhibit a consistent encoding profile, likely due to heterogeneity in response direction or magnitude across neural populations. Thus, the absence of a significant group-level effect in OFC may reflect divergent representational formats, rather than a lack of reward sensitivity.

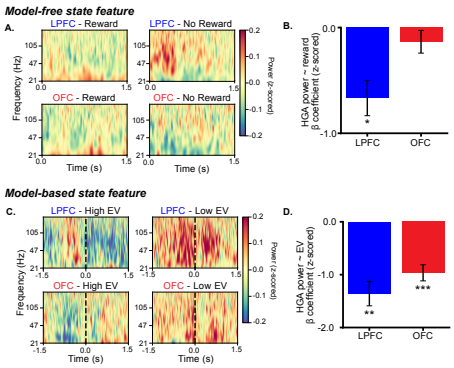

Figure 3: OFC and LPFC HGA consistently encode model-based value signals across participants. A. Normalized power for one exemplar participant (OFC= 10 electrodes; LPFC= 12 electrodes) separated by reward ($t = 0$ is choice/reward). B. Reward was significantly encoded in LPFC ($\beta = -0.70$, $z = -2.10$, $p < 0.05$) but not OFC ($\beta = -0.19$, $z = -1.11$, $p > 0.05$) HGA. C. Normalized power for one exemplar participant (OFC= 6 electrodes; LPFC= 5 electrodes) separated by EV ($t = 0$ is choice/reward). D. EV was significantly encoded in LPFC ($\beta = -1.33$, $z = 03.25$, $p < 0.01$) and OFC ($\beta = -1.031$, $z = -5.48$, $p < 0.001$) HGA, indicating a consistent neural representation of model-based value in both regions.

To test whether OFC and LPFC encode model-based state features consistently across participants, we replaced the reward regressor with expected value (EV) from the SA-RL model, controlling for relevant dimension and chosen features. Unlike reward, EV is internally computed through learning and reflects a participant's belief about expected outcomes. While EV and reward are correlated, using EV allows us to dissociate reactive outcome responses from features of proactive, belief-driven state representations. Because EV captures internal states shaped by learning and attentional strategy, it offers a more sensitive measure of how brain regions organize and maintain task-relevant information, and whether they share representational formats across individuals.

Both LPFC and OFC HGA demonstrated significant and consistent encoding of EV at the group level (LPFC: $\beta = -1.33$, $z = -3.25$, $p < 0.01$; OFC: $\beta = -1.031$, $z = -5.48$, $p < 0.001$; Fig. 3C/D) indicating homogeneity in the neural representation of model-based value estimates in these regions across participants. In contrast, two control regions, insula and anterior cingulate

cortex, did not exhibit the same consistency in EV encoding (Fig. S1), underscoring the specificity of representational similarity between OFC and LPFC in the context of this RL task.

The observed similarities in OFC and LPFC's HGA EV encoding suggest that, despite their different roles, these regions participate in a coordinated representational system that tracks belief-driven value signals necessary for maintaining adaptive state representations to guide flexible decision making (Cai & Padoa-Schioppa, 2014; Balewski et al., 2023; Rich & Wallis, 2016; Wilson et al., 2014; Schuck et al., 2016). This cross-region similarity suggests representational alignment at the neural level, pointing to a common computational format for encoding task-relevant information. By generating interpretable latent variables, this cognitive model-based approach sheds light on representations that are shared between regions, offering a framework to measure representational alignment at the neural level (Sucholutsky et al., 2024).

Parameterized Selective Attention Captures Individuals Differences in Local Features of Neural Representations: To assess individual differences in state representations across participants, we regressed the z-scored EV β coefficients against the SA-RL model's fitted selective attention (Φ; random effect = participant). We found selective attention selectively modulated LFPC EV encoding (LPFC: $\beta = -5.68$, $z = -2.19$, $p < 0.05$, Fig. 4A; OFC: $\beta = -1.23$, $z = -0.65$, $p > 0.05$, Fig. 4B). Participants who deployed similar strategies showed more similar LPFC representations of EV, indicating that subject-level computational alignment is mirrored in neural population activity. This finding highlights LPFC's role in directing attention to relevant information. Furthermore, we confirmed that this effect was not attributed to electrode placement (Fig. S2). By operationalizing latent selective attention, we reveal individual differences in neural state representations linked to task performance.

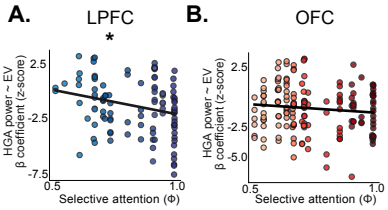

Figure 4: LPFC value signals are biased by attention. A. Significant relationship between selective attention (Φ) LPFC EV encoding ($\beta = -5.68$, $z = -2.19$, $p < 0.05$). B. No significant relationship between selective attention (Φ) and OFC EV encoding ($\beta = -1.23$, $z = -0.65$, $p > 0.05$).

Shared Neural Representations Exhibit Coordinated Temporal Profiles: Next, we investigated the temporal alignment of OFC and LPFC's EV-based state representations. Time-resolved multivariate regression was performed on each electrode (Fig 5B), and the time point corresponding with the strongest β coefficient for EV predicting HGA in each region was extracted. We found no significant difference in peak EV encoding between the regions ($D(159, 111) = 0.18$, $p > 0.05$; Fig. 5C). These findings suggest that OFC and LPFC exhibit temporally aligned EV encoding profiles, priming them for functional coordination (Enel et al., 2020).

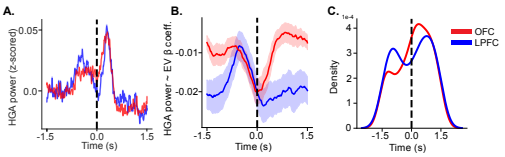

Figure 5: Similar temporal dynamics of OFC and LPFC EV encoding. A. HGA power across all OFC (red) and LPFC (blue) electrodes (shading = SEM, dashed line = choice/reward). B. Mean EV β coefficient across time for OFC (red) and LPFC (blue; shading = SEM). C. Peak EV encoding times for HGA in OFC (red) and LPFC (blue) showed no significant difference ($D(159, 111) = 0.18$, $p > 0.05$; dashed = choice/reward).

Parameterized Selective Attention Captures Individuals Differences in Inter-regional Features of Neural Representations: Next, to understand how the OFC and LPFC's independent EV-based representations are coordinated to support adaptive behavior, we focused on theta connectivity, given its role in inter-regional synchrony that facilitates sustained, selective attention (Fries, 2023; Landau et al., 2015). First, we examined the presence of theta oscillations (Fig. S3). Then, we measured directed LPFC-OFC theta connectivity using normalized PSI, comparing it to a null distribution (500 permutations) for high and low EV trials. PSI measures the slope of the phase difference between signals, capturing the direction and strength of phase coupling. Positive PSI values indicate information flow from one region to another, while negative values suggest the reverse direction. PSI was computed for high and low attention participants, based on SA-RL attention weights (Φ), to assess individual differences in LPFC-OFC connectivity based on attentional strategy.

A temporal shift in LPFC-OFC connectivity, linked to attention, was observed using a two-sample cluster test (500 permutations), with increased selective attention driving a shift from post- to pre-choice LPFC-OFC theta connectivity (Fig. 6A). The attention-modulated timing of LPFC-OFC coordination reflects its adaptive role in RL (Cai & Padoa-Schioppa, 2014), with neural mechanisms varying based on attentional state. In high attention individuals, LPFC-OFC connectivity peaks pre-choice, while in those with lower attention, it shifts post-choice, reflecting greater need for outcome-driven updating and attention switching in individuals with more distributed attention across state features.

Model-free attention splits (reaction time, performance) revealed LPFC-OFC theta connectivity (Fig. 6B), but lacked the temporal specificity of model-based attention. This finding highlights the utility of model-based neural analyses for revealing dynamics of shared neural representations in individuals using similar computational strategies for adaptive decision-making.

4 Conclusion

Leveraging cognitive modeling and iEEG to measure representational alignment, we identified shared computational strategies across individuals reflected in convergent neural representations across brain regions. Alignment was strongest among participants with similar cognitive strategies (Sucholutsky et al., 2024), suggesting that cognitive model-based metrics provide valuable insight into shared neural representations across biological systems. We operationalized neural representational alignment as the similarity in spectral and temporal properties of HGA EV encoding across brain regions and individuals. OFC and LPFC encoded model-based EV with comparable time–frequency signatures, suggesting a shared representational format across anatomically distinct regions. At the individual level, participants with similar model–derived attention strategies exhibited more similar LPFC encoding patterns and LPFC–OFC theta-band connectivity.

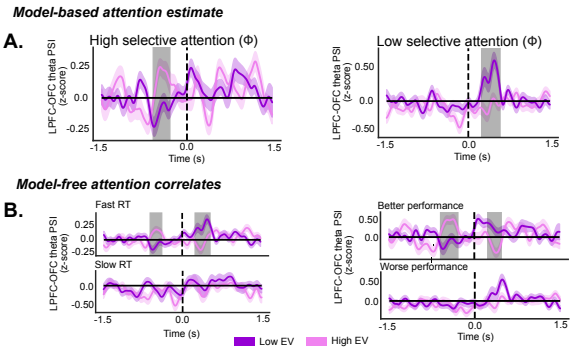

Figure 6: Model-based attention captures EV-based differences in LPFC-OFC connectivity. A. Participants were grouped into high (n=10) and low (n=11) attention based on SA-RL selective attention (Φ). Theta PSI was computed for high (light purple) and low (dark purple) EV trials. Greater attention shifted LPFC-OFC theta connectivity from post- to pre-choice (gray = significant cluster, $p < 0.05$). B. Median splits by reaction time (left) and performance (right) showed significant clusters ($p < 0.05$, gray), but lacked the temporal precision seen with model-based attention.

While this approach offers novel insights into representational alignment, several limitations remain. The SA-RL model isolates the role of selective attention in value learning but does not capture other cognitive processes, potentially oversimplifying real-world decision-making. Although spectral and temporal features provide a rich description of neural dynamics, our findings are correlational and do not establish causal links between cognitive strategies and neural encoding.

Nonetheless, our results show that parameters derived from cognitive models capture individual differences in behavioral strategies and predict corresponding variations in neural population dynamics. Using the SA-RL model, we discovered latent organizational principles that give rise to aligned representations across individuals and circuits. This integrative approach offers a framework for linking internal computational states with neural population activity, showing how shared cognitive strategies can lead to convergent representations across different brain regions and individuals. By identifying the neurocomputational principles behind aligned representations that relate to adaptive behavior in complex, real-world environments, we provide insights relevant for understanding and developing intelligent systems.

REFERENCES

Zuzanna Z. Balewski, Thomas W. Elston, Eric B. Knudsen, and Joni D. Wallis. Value dynamics affect choice preparation during decision-making. *Nature Neuroscience*, 26(9):1575–1583, September 2023. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-023-01407-3.

Xinying Cai and Camillo Padoa-Schioppa. Contributions of Orbitofrontal and Lateral Prefrontal Cortices to Economic Choice and the Good-to-Action Transformation. *Neuron*, 81(5):1140–1151, March 2014. ISSN 08966273. doi: 10.1016/j.neuron.2014.01.008.

Pierre Enel, Joni D Wallis, and Erin L Rich. Stable and dynamic representations of value in the prefrontal cortex. *eLife*, 9:e54313, July 2020. ISSN 2050-084X. doi: 10.7554/eLife.54313.

Pascal Fries. Rhythmic attentional scanning. *Neuron*, 111(7):954–970, April 2023. ISSN 08966273. doi: 10.1016/j.neuron.2023.02.015.

Ayelet Nina Landau, Helene Marianne Schreyer, Stan van Pelt, and Pascal Fries. Distributed Attention Is Implemented through Theta-Rhythmic Gamma Modulation. *Current Biology*, 25(17): 2332–2337, August 2015. ISSN 09609822. doi: 10.1016/j.cub.2015.07.048.

Yuan Chang Leong, Angela Radulescu, Reka Daniel, Vivian DeWoskin, and Yael Niv. Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, 93(2):451–463, January 2017. ISSN 08966273. doi: 10.1016/j.neuron.2016.12.040.

Christina Maher, Salman Qasim, Lizbeth Nuñez Martinez, Ignacio Saez, and Angela Radulescu. Intracranial recordings reveal neural encoding of attention-modulated reinforcement learning in humans. *Computational Cognitive Neuroscience*, August 2024.

Earl K Miller and Timothy J Buschman. Cortical circuits for the control of attention. *Current Opinion in Neurobiology*, 23(2):216–222, April 2013. ISSN 09594388. doi: 10.1016/j.conb.2012.11.011.

Yuval Nir, Lior Fisch, Roy Mukamel, Hagar Gelbard-Sagiv, Amos Arieli, Itzhak Fried, and Rafael Malach. Coupling between Neuronal Firing Rate, Gamma LFP, and BOLD fMRI Is Related to Interneuronal Correlations. *Current Biology*, 17(15):1275–1285, August 2007. ISSN 09609822. doi: 10.1016/j.cub.2007.06.066.

Yael Niv. Learning task-state representations. *Nature Neuroscience*, 22(10):1544–1553, October 2019. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-019-0470-8.

Yael Niv, Reka Daniel, Andra Geana, Samuel J. Gershman, Yuan Chang Leong, Angela Radulescu, and Robert C. Wilson. Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *The Journal of Neuroscience*, 35(21):8145–8157, May 2015. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.2978-14.2015.

Erin L Rich and Jonathan D Wallis. Decoding subjective decisions from orbitofrontal cortex. *Nature Neuroscience*, 19(7):973–980, July 2016. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.4320.

Erin L. Rich and Joni D. Wallis. Spatiotemporal dynamics of information encoding revealed in orbitofrontal high-gamma. *Nature Communications*, 8(1):1139, October 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-01253-5.

Ignacio Saez, Jack Lin, Arjen Stolk, Edward Chang, Josef Parvizi, Gerwin Schalk, Robert T. Knight, and Ming Hsu. Encoding of Multiple Reward-Related Computations in Transient and Sustained High-Frequency Activity in Human OFC. *Current Biology*, 28(18):2889–2899.e3, September 2018. ISSN 09609822. doi: 10.1016/j.cub.2018.07.045.

Nicolas W. Schuck, Ming Bo Cai, Robert C. Wilson, and Yael Niv. Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron*, 91(6):1402–1412, September 2016. ISSN 08966273. doi: 10.1016/j.neuron.2016.08.019.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J. Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *arXiv preprint*, arXiv:2310.13018, November 2024. doi: 10.48550/arXiv.2310.13018.

Robert C. Wilson and Yael Niv. Inferring Relevance in a Changing World. *Frontiers in Human Neuroscience*, 5, 2012. ISSN 1662-5161. doi: 10.3389/fnhum.2011.00189.

Robert C. Wilson, Yuji K. Takahashi, Geoffrey Schoenbaum, and Yael Niv. Orbitofrontal Cortex as a Cognitive Map of Task Space. *Neuron*, 81(2):267–279, January 2014. ISSN 08966273. doi: 10.1016/j.neuron.2013.11.005.

A APPENDIX

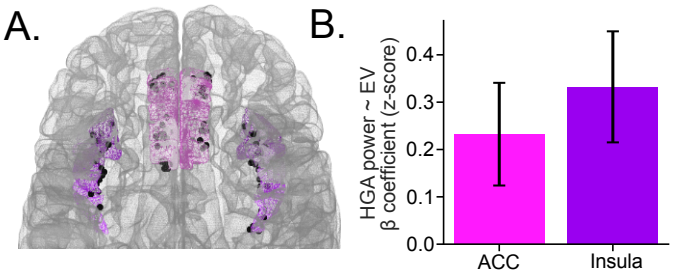

Figure S1: ACC and insula HGA do not consistently encode model-based value signals across participants. A. Electrode locations (black) in the ACC (pink, 123 electrodes) and insula (purple, 101 electrodes) across 21 iEEG participants. B. Using the same regression approach as in Fig. 3D, we tested whether ACC and insula trial-averaged HGA consistently encoded EV of choice (controlling for relevant dimension and chosen features). As expected, neither region significantly encoded EV (ACC: ($\beta = 0.13$, $z = 0.74$, $p > 0.05$); insula: ($\beta = 0.30$, $z = 1.60$, $p > 0.05$), suggesting that, unlike OFC and LPFC, ACC and insula do not exhibit a consistent representation of model-based value.

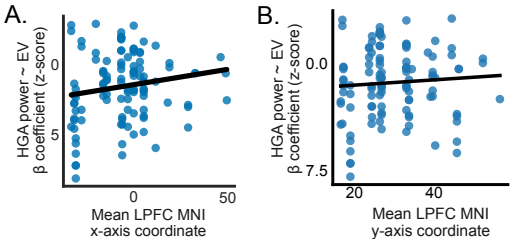

Figure S2: LPFC's attention-modulated value signals are not driven by variability in electrode placement. To rule out the influence of individual differences in electrode placement, we repeated the analysis from Fig. 4A, this time regressing EV β coefficient z-scores against participants' A. average MNI x-axis coordinate and B. average MNI y-axis coordinate for LPFC electrodes, with participant as a random effect. The analysis confirmed that electrode placement does not account for the observed effect (MNI x-axis: $\beta = 0.01$, $z = 0.70$, $p > 0.05$; MNI y-axis: $\beta = 0.02$, $z = 0.97$, $p > 0.05$).

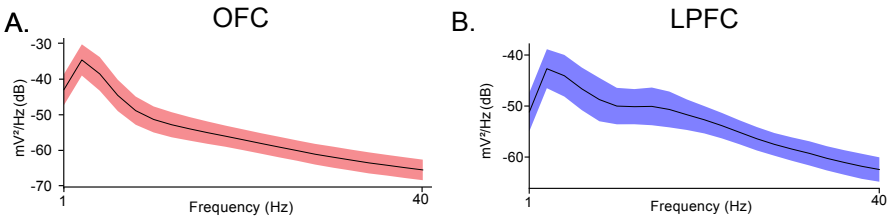

Figure S3: Theta oscillations detected in OFC and LPFC recordings. A. Power spectral density (PSD) for OFC (5 electrodes; shading = SEM) showing a peak in the theta range (2-8 Hz) in an exemplar patient. B. Power spectral density (PSD) for LPFC (9 electrodes; shading = SEM) also revealing a peak in the theta range, indicating the presence of oscillations in this frequency range.