Extended Abstract Track

Factorized Prefrontal Geometry of Goal and Uncertainty Explains Flexible yet Stable Human Goal Pursuit

Anonymous Authors

Editors: List of editors' names

Abstract

A central challenge for adaptive agents is achieving behavioral flexibility without losing stability, especially during goal-directed learning in uncertain environments. Flexibility enables rapid adjustment when goal changes, while stability prevents overreaction to noisy outcomes— these properties often trade off. To examine how the brain resolves this dilemma, we combined reinforcement learning simulations, human behavior, and fMRI study during a sequential decision task with varying goals and uncertainty. Simulations showed that model-free agents suffered a flexibility–stability trade-off, while model-based agents were flexible but with a range of stability. Human participants, in contrast, displayed both high flexibility and high stability. fMRI analyses revealed the underlying representational mechanism: goals and uncertainty were encoded as partly independent dimensions in lateral prefrontal and orbitofrontal cortex. Importantly, the separability and robustness of goal representations in these regions correlated with individual behavioral flexibility and stability, pointing to a geometric account of robust goal pursuit beyond value-learning models.

Keywords: Representational geometry; cognitive flexibility; stability; uncertainty; LPFC; OFC; MVPA; fMRI; reinforcement learning.

1. Introduction

Robust goal pursuit in uncertain environments requires balancing two opposing demands: flexibility, the ability to adapt when goals change, and stability, the ability to resist noise and avoid erratic behavior. Overemphasizing one typically compromises the other, creating a long-recognized stability–flexibility dilemma (Goschke, 2013; Hommel, 2015; Dreisbach and Fröber, 2019; Musslick and Cohen, 2021; Qiao et al., 2023).

The prefrontal cortex (PFC) plays an important role in this problem. It encodes task-relevant information, adapts its representations to goals, and tracks environmental uncertainty through regions such as the lateral PFC (LPFC) and orbitofrontal cortex (OFC) (Huettel et al., 2005; Soltani and Izquierdo, 2019; Hsu et al., 2005; Soltani and Koechlin, 2022). Yet theories of neural coding suggest a tension: high-dimensional mixed selectivity supports flexible readouts (Rigotti et al., 2013; Tang et al., 2019; Sheng et al., 2022), while low-dimensional abstraction promotes stable performance (Mack et al., 2020; Bernardi et al., 2020; Flesch et al., 2022). How the brain reconciles these conflicting representational demands during sequential, goal-directed learning under uncertainty remains unclear (Fusi et al., 2016; Badre et al., 2021; Jazayeri and Ostojic, 2021; Chung and Abbott, 2021).

In this work, we approach the problem by combining behavioral analysis, reinforcement learning models, and fMRI-based geometry of neural codes. This framework allows us to ask whether the PFC organizes goal and uncertainty signals in a way that preserves goal stability

while enabling adaptive control under noise. More broadly, it provides a representational perspective that can address aspects of robust goal pursuit not explained by value-learning theories alone.

2. Methods

Task Twenty adults performed a two-stage Markov decision task with manipulations of goal (specific: only one designated coin color rewarded; non-specific: any color rewarded) and uncertainty (low: 0.9/0.1 transitions; high: 0.5/0.5; Fig. 1a–b). This design isolated the effects of goal specificity and environmental predictability. Human behavior was compared with model-based (MB) and model-free (MF) reinforcement learning agents simulated on the identical block schedules. See Appendix A for details.

Behavioral metrics We quantified behavioral flexibility (switching choices when goals changed; *choice versatility*), stability (repeating choices when goals stayed constant despite noise; *choice consistency*), and performance (trial-wise optimality relative to an oracle agent; *choice optimality*). These metrics allowed us to test whether humans and RL agents achieve robust goal pursuit by balancing adaptation with consistency.

ROI-based fMRI analysis and neural metrics BOLD activity was analyzed using multivoxel pattern analysis (MVPA) with linear SVMs in eight predefined anatomical ROIs (vlPFC, dlPFC, OFC, ACC, preSMA, V1, HPC, vStr; Fig. 5). We assessed four neural properties: (i) decoding accuracy for goal and uncertainty, (ii) *shattering dimensionality* (SD) (Rigotti et al., 2013), which tested the linear separability of all dichotomies based on the combination of the two variables, (iii) *cross-condition generalization performance* (CCGP) (Bernardi et al., 2020), which measured whether a linear decoder trained at one context level generalized to the other, and (iv) *parallelism score* (PS), the cosine similarity between coding directions estimated separately in different contexts.

3. Results

Robust human behavior emerges only under explicit goal pursuit. Value-based decision-making theory predicts that larger action-value differences facilitate optimal choices, whereas uncertainty reduces these differences (Fig. 1c). Consistent with this, human participants exhibited reduced behavioral performance when no specific goal was provided; however, when pursuing a specific goal, their performance remained robust against uncertainty (Fig. 1d), indicating that stability emerges during goal-directed learning beyond value-based accounts.

High flexibility and stability in human goal-directed learning To examine this robustness, we quantified flexibility (adapting to changing goals) and stability (resilience under uncertainty; Fig. 1e) and compared humans with simulated MB and MF agents. Notably, MF agents exhibited a flexibility–stability trade-off, whereas MB agents and humans did not. Flexibility and stability both supported optimality in humans and MB agents. Strikingly, humans achieved the highest levels of both flexibility and stability. These findings provided strong motivation to investigate how humans overcome the trade-off. We therefore probed the neural code supporting this profile.

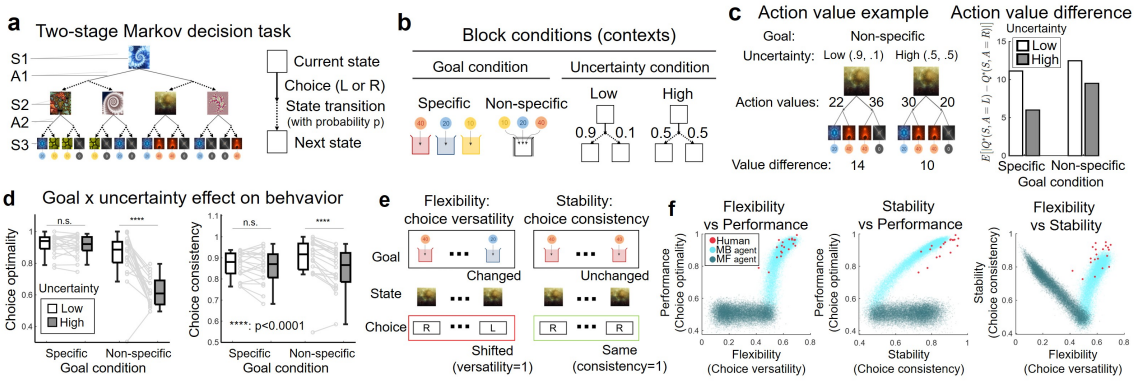

Figure 1: (a) Task structure. (b) Goal conditions: specific (colored box) vs. non-specific (white box). (c) Action-value differences under low vs. high uncertainty. (d) Effect of goal and uncertainty on performance ($n = 20$). (e) Definitions of behavioral flexibility and stability. (f) Human vs. simulated MB/MF agents (20k different parameter sets).

Uncertainty is encoded only during goal pursuit, and neural dimensionality expands with specific goals. We next asked where goal and uncertainty are represented. Goal decoding was significant in the vlPFC, dlPFC, OFC, ACC, preSMA, and V1 (Fig. 2a–b), while uncertainty was reliably encoded only in the vlPFC, dlPFC, and OFC, only under specific goals (Fig. 2c–d). Thus, uncertainty sensitivity was conditional on goal pursuit. Moreover, neural effective dimensionality increased in the vlPFC, dlPFC, OFC, and ACC during specific goals (Fig. 6), suggesting that PFC regions expand into higher-dimensional codes to support robust goal pursuit.

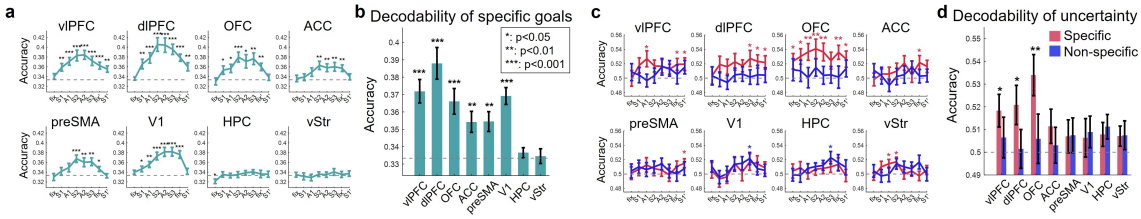

Figure 2: (a, c) Goal/uncertainty decoding across trial events (dashed line: chance level). Error bars = SEM. (b, d) Average goal/uncertainty decoding accuracy.

Factorized embedding of goal and uncertainty in the LPFC To test how goals remain stable under uncertainty, we analyzed representational geometry, considering three regimes (Fig. 3a): *compression* (only one variable represented, stable but limited), *factorized mixing* (variables linearly independent, generalizable across each other), and *nonlinear mixing* (axes rotate, higher dimensionality but poor generalization). The nature of possible linear separations depends on the complexity of the underlying neural embedding (Vapnik and Chervonenkis, 2015; Abu-Mostafa, 1989). Using shattering analysis of fMRI data, we tested the separability of all possible dichotomies (Fig. 3a) and grouped them into four categories (goal, uncertainty, linear, nonlinear; Fig. 7) to derive region-specific shattering profiles (Fig. 3b). The vlPFC, dlPFC, and OFC showed factorized profiles, jointly encod-

ing goal and uncertainty while maintaining stable embeddings. The ACC, preSMA, and V1 exhibited compression, encoding goals alone; consistent with decoding results (Fig. 2). Neural goal separability predicted individual flexibility, stability, and optimality, with the vlPFC and dlPFC explaining variance in all three measures (Fig. 3c).

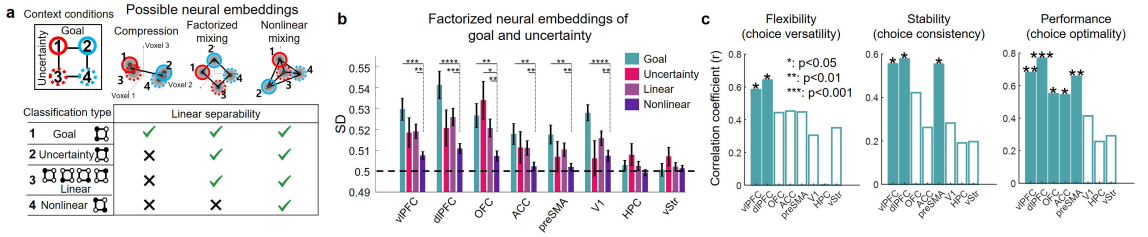

Figure 3: (a) Hypothetical embedding regimes. (b) Shattering dimensionality (SD) across dichotomy types. (c) Correlation coefficients between goal SD and behavior.

Neurally stable goal embedding in LPFC guides stably flexible learning Goal representations remained robust across uncertainty: CCGP and SD were comparable between conditions (Fig. 4a). Despite representing uncertainty, the vlPFC, dlPFC, and OFC preserved stable goal axes, and higher robustness in the vlPFC and dlPFC correlated with behavioral flexibility, stability, and overall performance (Fig. 4b). PS further confirmed aligned goal-encoding directions across uncertainty in the vlPFC, dlPFC, OFC, and preSMA (Fig. 4c-d), supporting minimal reorientation. Together, CCGP and PS demonstrate that PFC maintains invariant goal readouts, enabling stably flexible behavior.

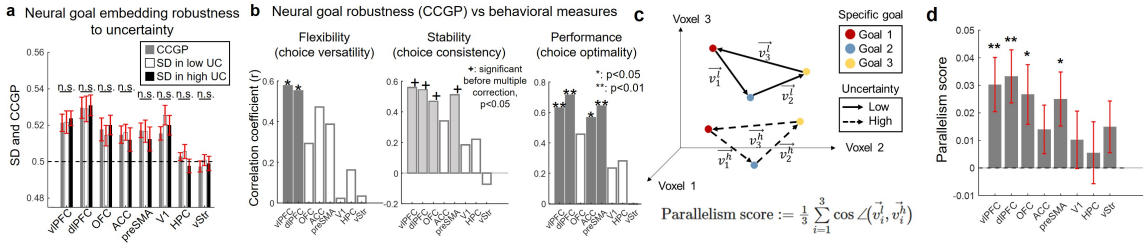

Figure 4: (a) Goal CCGP and SD across uncertainty levels. (b) Correlations between goal CCGP and behavioral measures. (c) PS: schematic and definition. (d) ROI-wise PS results.

4. Discussion

Prior works located goal- and uncertainty-related signals in PFC but rarely addressed how stable performance is maintained under uncertainty; our data indicate that factorized goal x uncertainty codes in the LPFC and OFC mitigate the flexibility–stability dilemma and offer a tractable representational account that complements value-learning models. By independently encoding uncertainty (for strategy selection) and goal (for action selection), LPFC can organize a stable hierarchy of strategy and action across contexts and stages; an aspect left open by many single-context studies. Looking forward, testing agents with factorized goal–uncertainty embeddings and extending to longer timescales and richer tasks will clarify when this geometry reproduces human-like robustness.

References

Yaser S. Abu-Mostafa. The Vapnik-Chervonenkis Dimension: Information versus Complexity in Learning. *Neural Computation*, 1(3):312–317, September 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.3.312.

David Badre, Apoorva Bhandari, Haley Keglovits, and Atsushi Kikumoto. The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, 38: 20–28, April 2021. ISSN 2352-1546. doi: 10.1016/j.cobeha.2020.07.002.

Silvia Bernardi, Marcus K. Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C. Daniel Salzman. The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, 183(4):954–967.e21, November 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.09.031.

SueYeon Chung and L. F. Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*, 70: 137–144, October 2021. ISSN 0959-4388. doi: 10.1016/j.conb.2021.10.010.

Gesine Dreisbach and Kerstin Fröber. On How to Be Flexible (or Not): Modulation of the Stability-Flexibility Balance. *Current Directions in Psychological Science*, 28(1):3–9, February 2019. ISSN 0963-7214. doi: 10.1177/0963721418800030.

Simon B. Eickhoff, Tomas Paus, Svenja Caspers, Marie-Helene Grosbras, Alan C. Evans, Karl Zilles, and Katrin Amunts. Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. *NeuroImage*, 36(3):511–521, July 2007. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2007.03.060.

Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, 110(7):1258–1270.e11, April 2022. ISSN 0896-6273. doi: 10.1016/j.neuron.2022.01.005.

Stefano Fusi, Earl K Miller, and Mattia Rigotti. Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37:66–74, April 2016. ISSN 0959-4388. doi: 10.1016/j.conb.2016.01.010.

Thomas Goschke. Volition in Action: Intentions, Control Dilemmas, and the Dynamic Regulation of Cognitive Control. February 2013. doi: 10.7551/mitpress/9780262018555.003.0024.

Bernhard Hommel. Chapter Two - Between Persistence and Flexibility: The Yin and Yang of Action Control. In Andrew J. Elliot, editor, *Advances in Motivation Science*, volume 2, pages 33–67. Elsevier, January 2015. doi: 10.1016/bs.adms.2015.04.003.

Ming Hsu, Meghana Bhatt, Ralph Adolphs, Daniel Tranel, and Colin F. Camerer. Neural Systems Responding to Degrees of Uncertainty in Human Decision-Making. *Science*, 310 (5754):1680–1683, December 2005. doi: 10.1126/science.1115327.

Scott A. Huettel, Allen W. Song, and Gregory McCarthy. Decisions under Uncertainty: Probabilistic Context Influences Activation of Prefrontal and Parietal Cortices. *Journal of Neuroscience*, 25(13):3304–3311, March 2005. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.5070-04.2005.

Mehrdad Jazayeri and Srdjan Ostojic. Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Current Opinion in Neurobiology*, 70:113–120, October 2021. ISSN 0959-4388. doi: 10.1016/j.conb.2021.08.002.

Michael L. Mack, Alison R. Preston, and Bradley C. Love. Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, 11(1):46, January 2020. ISSN 2041-1723. doi: 10.1038/s41467-019-13930-8.

Sebastian Musslick and Jonathan D. Cohen. Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, 25(9):757–775, September 2021. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2021.06.001.

Lei Qiao, Lijie Zhang, and Antao Chen. Control dilemma: Evidence of the stability–flexibility trade-off. *International Journal of Psychophysiology*, 191:29–41, September 2023. ISSN 0167-8760. doi: 10.1016/j.ijpsycho.2023.07.002.

Mattia Rigotti, Omri Barak, Melissa R. Warden, Xiao-Jing Wang, Nathaniel D. Daw, Earl K. Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, May 2013. ISSN 1476-4687. doi: 10.1038/nature12160.

Edmund T. Rolls, Chu-Chung Huang, Ching-Po Lin, Jianfeng Feng, and Marc Joliot. Automated anatomical labelling atlas 3. *NeuroImage*, 206:116189, February 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2019.116189.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pages 606–610, September 2007.

Jintao Sheng, Liang Zhang, Chuqi Liu, Jing Liu, Junjiao Feng, Yu Zhou, Huinan Hu, and Gui Xue. Higher-dimensional neural representations predict better episodic memory. *Science Advances*, 8(16):eabm3829, April 2022. doi: 10.1126/sciadv.abm3829.

Alireza Soltani and Alicia Izquierdo. Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, 20(10):635–644, October 2019. ISSN 1471-0048. doi: 10.1038/s41583-019-0180-y.

Alireza Soltani and Etienne Koechlin. Computational models of adaptive behavior and prefrontal cortex. *Neuropsychopharmacology*, 47(1):58–71, January 2022. ISSN 1740-634X. doi: 10.1038/s41386-021-01123-1.

Evelyn Tang, Marcelo G. Mattar, Chad Giusti, David M. Lydon-Staley, Sharon L. Thompson-Schill, and Danielle S. Bassett. Effective learning is accompanied by high-dimensional and efficient representations of neural activity. *Nature Neuroscience*, 22(6):1000–1009, June 2019. ISSN 1546-1726. doi: 10.1038/s41593-019-0400-9.

V. N. Vapnik and A. Ya. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. In Vladimir Vovk, Harris Papadopoulos, and Alexander Gammerman, editors, *Measures of Complexity: Festschrift for Alexey Chervonenkis*, pages 11–30. Springer International Publishing, Cham, 2015. ISBN 978-3-319-21852-6. doi: 10.1007/978-3-319-21852-6_3.

Appendix A. Methods (expanded)

Task and participants We analyzed the behavioral and fMRI dataset from a two-stage Markov decision task with uncertainty-changing blocks. Twenty-two right-handed adults were recruited; two were excluded (one for invariant Stage-1 choices, one for sub-chance performance), yielding $n = 20$. Each trial comprised two left/right choices to reach a colored coin outcome; coins (red/yellow/blue) had participant-specific values. After 100 pretraining trials with 0.5/0.5 transitions, participants completed five scanning sessions (about 80 trials each, total 400). Block contexts crossed goal (specific: only the cued color pays; non-specific: any color pays) with uncertainty (low: 0.9/0.1; high: 0.5/0.5) in short alternating blocks. Choices timed out at 4s; ITIs were 1–4s; outcomes were displayed for 2s. Participants were told that goals and transition probabilities could change but not the numeric probabilities.

Behavioral measures We summarized behavior with three trial-level metrics and averaged them per participant. *Choice optimality* scored 1 when the chosen action matched an oracle agent's optimal action for that state and context (ties excluded). *Choice versatility* (flexibility) scored 1 when, at the same state, the current choice switched relative to the previous trial only when the goal changed. *Choice consistency* (stability) scored 1 when, at the same state, the current choice repeated the previous choice only when the goal stayed the same, indexing resistance to noise-induced switching.

Simulated agents We simulated two reinforcement-learning agents under the exact block schedules used by humans (same trial counts, goal sequences, and uncertainty switches) to benchmark flexibility, stability, and performance. The model-based (MB) agent combined (i) *FORWARD* learning of transition dynamics via a state-prediction error (SPE) and (ii) *BACKWARD* planning that re-evaluates action values when the trial goal changes. Concretely, with transition matrix $T(s, a, s')$ and terminal rewards $r(s) = R$ for goal states (0 otherwise), the MB values were computed by dynamic programming

$$Q^{\mathrm{MB}}(s, a) = \sum_{s'} T(s, a, s') \Big[r(s') + \gamma \max_{a'} Q^{\mathrm{MB}}(s', a') \Big], \qquad \gamma = 1,$$

and the model was updated online using

$$\delta_{\mathrm{SPE}} = 1 - T(s, a, s'), \quad \Delta T(s, a, s') = \eta \, \delta_{\mathrm{SPE}},$$

with η the SPE learning rate. The model-free (MF) agent used SARSA with a reward-prediction error (RPE)

$$\delta_{\mathrm{RPE}} = r(s') + \gamma \, Q^{\mathrm{MF}}(s', a') - Q^{\mathrm{MF}}(s, a), \quad \Delta Q^{\mathrm{MF}}(s, a) = \alpha \, \delta_{\mathrm{RPE}},$$

where α is the value-learning rate and $\gamma = 1$. Both agents chose stochastically via a softmax policy

$$\Pr(a \mid s) = \frac{\exp\{\tau \, Q(s, a)\}}{\sum_{a'} \exp\{\tau \, Q(s, a')\}},$$

with inverse temperature τ (decision noise/exploitation). For each of the 20 human schedules we drew 1,000 random parameter sets per model (yielding 20,000 MB and 20,000 MF agents across schedules); each model had two free parameters: a learning rate (η or α) and τ. Agent behavior was scored with the same trial-wise metrics as humans: *choice optimality*, *choice versatility*, and *choice consistency*.

fMRI acquisition and ROIs Scanning used a 3T Siemens Trio (32-channel coil). Structural images: MPRAGE (TR = 1,500 ms; TE = 2.63 ms; flip 10°; 1 mm isotropic). Functional images: EPI (45 slices tilted 30° from AC–PC; TR = 2,800 ms; TE = 30 ms; flip 80°; 3 mm isotropic). Preprocessing in SPM8 included slice-timing correction, realignment, coregistration, and normalization to MNI152. For MVPA, voxel time series were detrended and z-scored within session. Eight bilateral ROIs were analyzed (Fig. 5): the ventrolateral prefrontal cortex (vlPFC), dorsolateral prefrontal cortex (dlPFC), orbitofrontal cortex (OFC), anterior cingulate cortex (ACC), pre-supplementary motor area (preSMA), primary visual cortex (V1), hippocampus (HPC), and ventral striatum (vStr). ROIs were defined from

AAL3, except preSMA (JuBrain); vlPFC: triangular inferior frontal gyrus; dlPFC: middle frontal gyrus; OFC: inferior, middle, superior orbital gyri + rectal gyrus; ACC: pregenual + supracallosal. Hemispheres were averaged; preliminary checks confirmed voxel-count differences did not drive results.

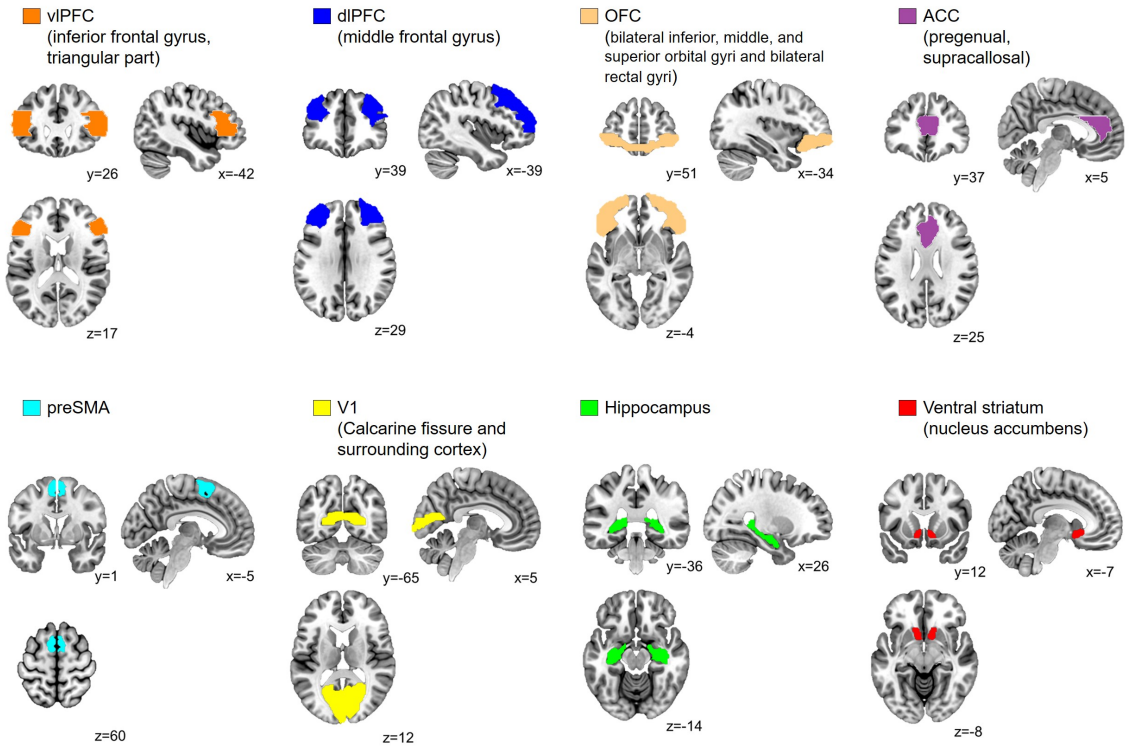

Figure 5: Anatomically defined ROIs. Binary masks for the eight ROIs—ventrolateral prefrontal cortex (vlPFC), dorsolateral PFC (dlPFC), orbitofrontal cortex (OFC), anterior cingulate cortex (ACC), pre-supplementary motor area (preSMA), primary visual cortex (V1), hippocampus (HPC), and ventral striatum (vStr)—are overlaid bilaterally on the MNI152 T1-weighted template (coronal, sagittal, and axial views). All masks were taken from the AAL3 atlas (Rolls et al., 2020) except the preSMA, which was obtained from the JuBrain Anatomy Toolbox (Eickhoff et al., 2007).

Unsupervised dimensionality For each ROI, goal condition (specific vs. non-specific), and uncertainty level (low vs. high), we quantified neural dimensionality with the *effective rank* of the PCA eigenspectrum (Roy and Vetterli, 2007). Trial-locked multivoxel patterns were extracted at S1, A1, S2, A2, S3, and the post-trial fixation (fix′); within each session voxel time series were detrended and z-scored. For a given condition we stacked all event-locked patterns, computed the voxelwise covariance, and performed PCA. Let $\{\lambda_i\}$ denote the non-negative eigenvalues (variance explained) of this covariance; we formed normalized weights $p_i = \lambda_i / \sum_j \lambda_j$ and the spectral entropy $H = -\sum_i p_i \log p_i$ (natural log). The effective rank is $\mathrm{ER} = \exp(H)$, which ranges from 1 (all variance in a single component) to the number of components (uniform spectrum); higher ER indicates greater usable dimensionality. ER was computed separately per event and then averaged across events

and hemispheres to yield one value per ROI per participant. Group effects of goal, uncertainty, and their interaction were tested with a two-way repeated-measures ANOVA per ROI (Fig. 6).

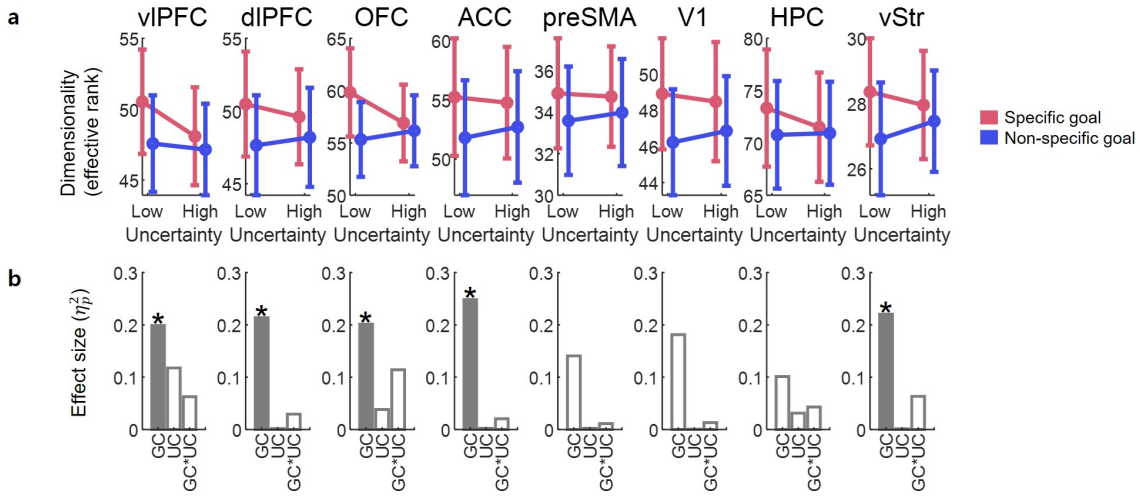

Figure 6: Effect of goal and uncertainty condition on the neural dimensionality of brain regions. (a) The neural dimensionality of brain regions measured under different goal and uncertainty conditions. The dimensionality was quantified using the effective rank measure (Roy and Vetterli, 2007). This measure involves performing PCA on the BOLD signal from each brain region to obtain the eigenspectrum, calculating its entropy H, and then exponentiating the result. As defined by the previous work (Roy and Vetterli, 2007), the effective rank $ER = \exp(H)$. Error bars represent the standard error across participants. Effective rank values were averaged across the left and right hemispheres. (b) The effect sizes obtained from a two-way repeated measures ANOVA. GC represents the goal condition, UC represents the uncertainty condition, and GC*UC denotes the interaction between these two conditions. Only significant effects are shown with filled bars (*: $p < 0.05$).

Decoding pipeline (events and cross-validation) All multivoxel analyses used linear SVMs trained and tested within each ROI with leave-one-session-out validation (five folds). Event-locked patterns were extracted at eight timestamps per trial: Fix1, S1, A1, Fix2, S2, A2, Fix3, S3. To respect hemodynamic lag and sequence timing, the volume immediately after an event indexed its response; because choices were followed by fixation cues, A1/A2 responses were taken from volumes after Fix2/Fix3. We decoded goals (3-way), uncertainty (binary), and dichotomies (see SD below) at each informative event; the initial fixation (Fix1) was excluded from SD, while the subsequent trial's fixation (fix') was used to capture residual post-outcome signals. Class imbalance was controlled by 100 rounds of undersampling with different seeds; reported accuracies are means across rounds and folds. Chance levels were 1/3 (goal) and 0.5 (binary tasks). Shuffled-label controls confirmed all metrics fell to chance.

Shattering dimensionality (SD): procedure and geometric rationale To diagnose representational geometry, we considered the six goal × uncertainty classes (3 goals × 2

uncertainty levels) and evaluated all unique binary labelings (31 dichotomies). Dichotomies were grouped into four types: goal (three one-vs-rest goal splits), uncertainty (low vs high across goals), linear (dichotomies linearly separable under a generic linear mixing of goal and uncertainty), and nonlinear (nine dichotomies empirically not linearly separable under that model). For each ROI and event (S1–fix'), we trained SVMs on every dichotomy and defined SD as the mean test accuracy within each type (chance level: 0.5). Predicted profiles: *compression* (high SD for only one variable), *factorized mixing* (high goal/uncertainty/linear SD; lower nonlinear SD), and *nonlinear mixing* (high SD for all types). This taxonomy links linear readout availability to embedding structure and generalization.

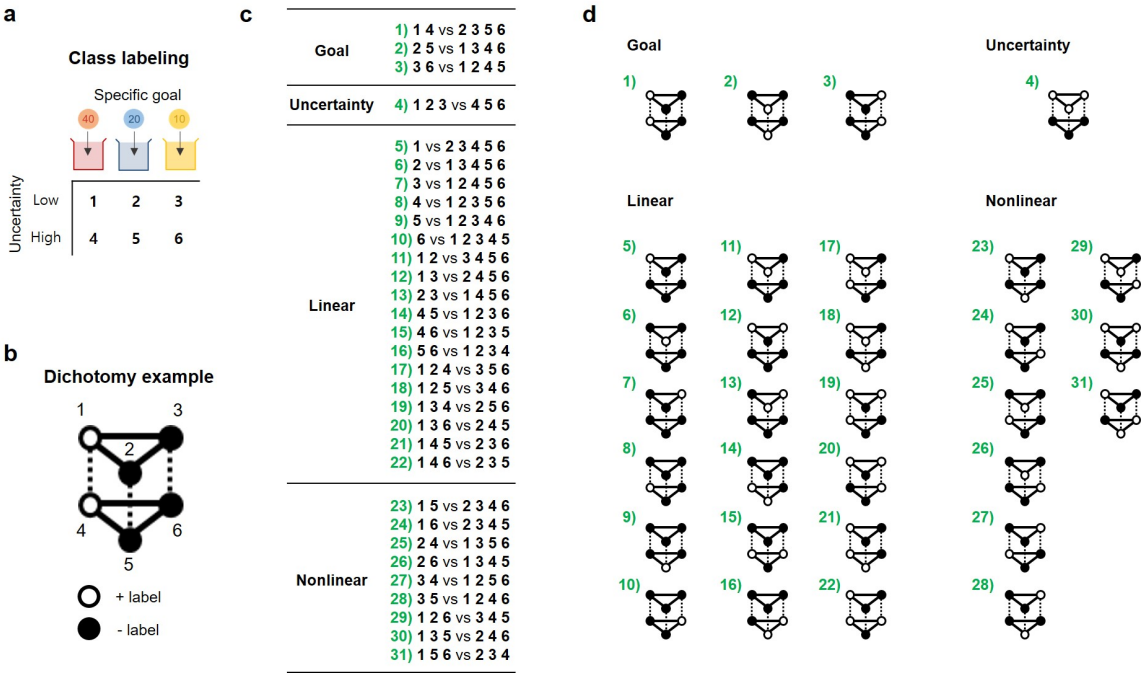

Figure 7: Categorization of all dichotomies. (a) The six classes determined by the specific goal and uncertainty conditions. (b) A schematic of a dichotomy based on the six classes. (c) The four categories (goal, uncertainty, linear, and nonlinear) for all dichotomies. There are 2^6 ways of binary labeling with the six classes. The actual number of dichotomies reduces to $\frac{2^6-2}{2}$ by excepting the two cases of all positive or negative labeling and removing half of the duplicated cases due to the symmetry of binary labeling. (d) Visualization of all categorized dichotomies.

Cross-condition generalization (CCGP) and parallelism CCGP tests whether a single goal readout transfers across uncertainty. For each ROI/event, we trained three goal dichotomies (e.g., red vs {blue,yellow}) in low uncertainty and tested in high, and vice versa; CCGP is the mean of the six cross-directions. Comparing CCGP to within-condition SD quantifies context sensitivity: large CCGP implies an uncertainty-invariant goal axis. A complementary parallelism score (PS) assessed alignment of coding directions: within each uncertainty level we formed mean patterns per goal and computed vectors between goal pairs; PS is the cosine similarity between corresponding low- vs high-uncertainty vectors,

averaged across pairs, events, and sessions (with undersampling for class balance). Positive PS indicates minimal axis rotation across uncertainty.

Statistics Participant-level accuracies (or ranks) were averaged across events as specified, yielding one value per ROI per metric. Hypothesis-driven tests were two-tailed at $\alpha = 0.05$ and conducted per predefined ROI without family-wise correction (no exploratory search across ROIs). Where multiple event types or SD categories were compared within an ROI, planned comparisons are reported without additional correction. Exploratory brain–behavior correlations were corrected across the eight ROIs using Benjamini–Hochberg FDR ($q = 0.05$); adjusted q values are reported. Chance levels were 1/3 (3-way goal) and 0.5 (binary tasks).