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

Yoondo Sung[1], Mattia Rigotti[2], Sang Wan Lee*[1,3,4,5,6]

[1]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST). [2]IBM Research. [3]Center for Neuroscience-inspired AI, KAIST. [4]Department of Brain and Cognitive Sciences, KAIST. [5]Kim Jaechul Graduate School of AI, KAIST. [6]Graduate School of Data Science, KAIST.

* Corresponding author | Contact: ydsung@kaist.ac.kr

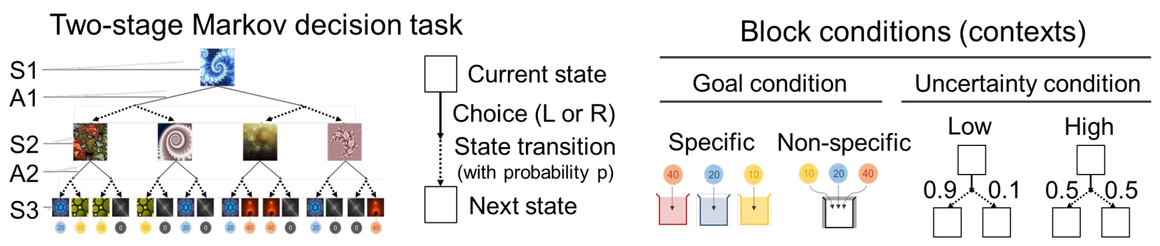
Question: behavioral stability-flexibility dilemma during goal-directed learning – how does the brain solve this problem?

Behavioral findings: human's stable goal pursuit behavior – 1) uncertainty-robust performance, 2) higher flex. & stab. than RL agents

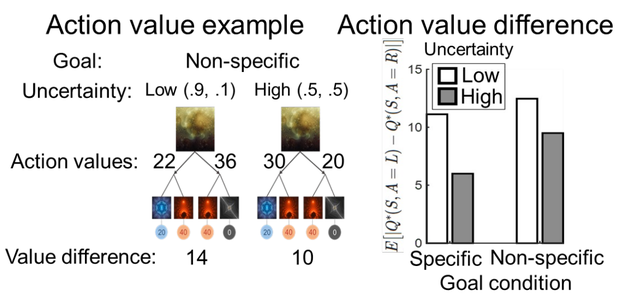
Neural findings: factorized embedding of goal and uncertainty in the LPFC ∝ flexible adaptation to goal change & stable goal pursuit

1 Behavioral stability-flexibility dilemma during goal-directed learning

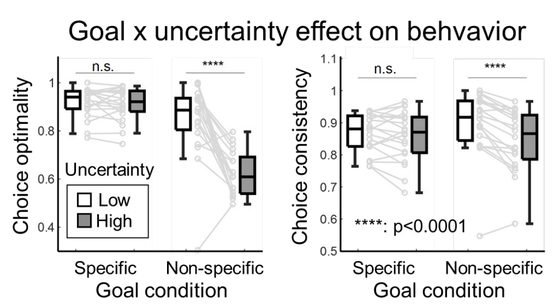
- Markov decision task with uncertainty and goal changes [1]



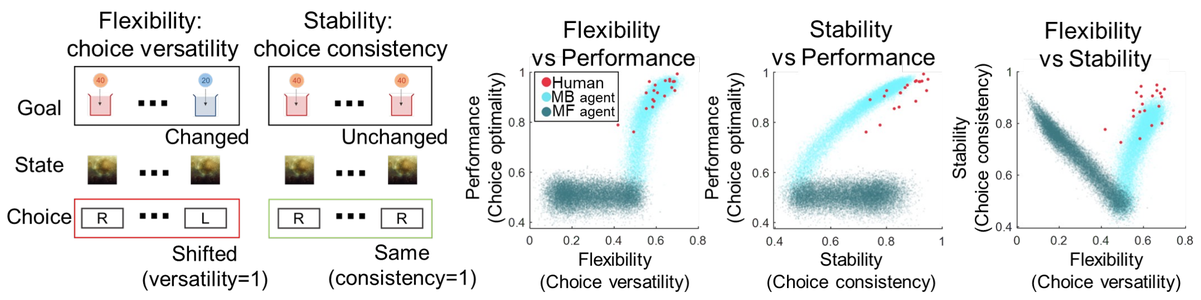
- Uncertainty hinders value-based decision making.



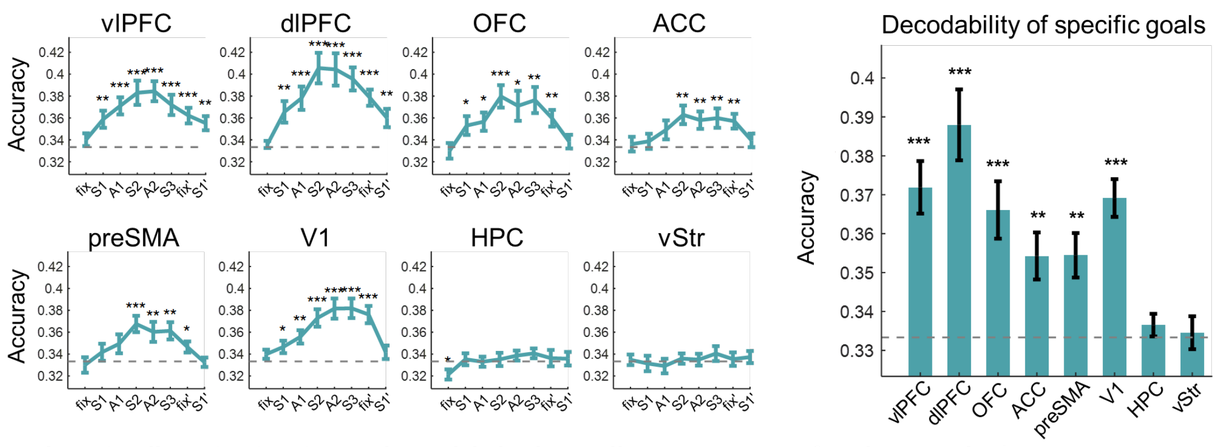
- Robust human behavior emerges only under explicit goal pursuit.



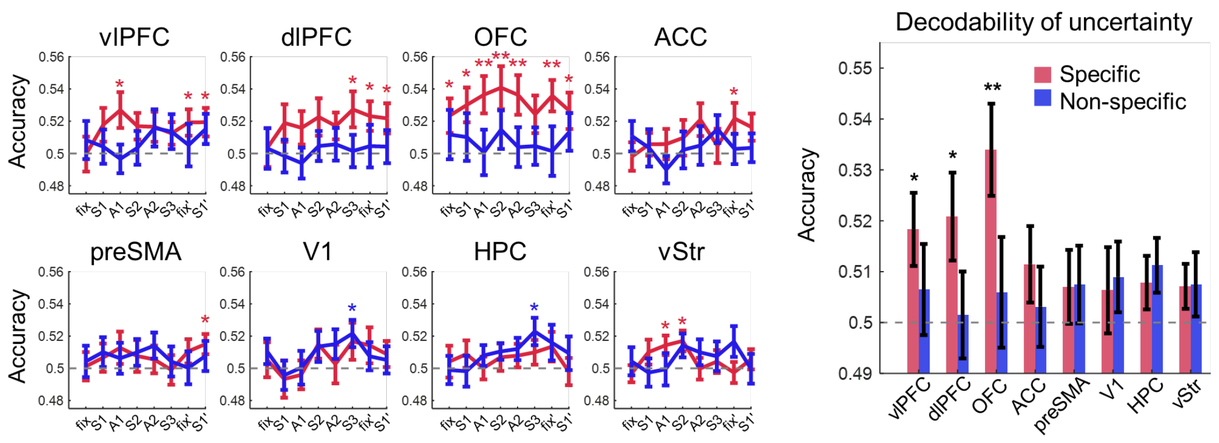
- High flexibility and stability in human goal-directed learning



2 Evidence of goal and uncertainty representation in PFC during goal-directed learning



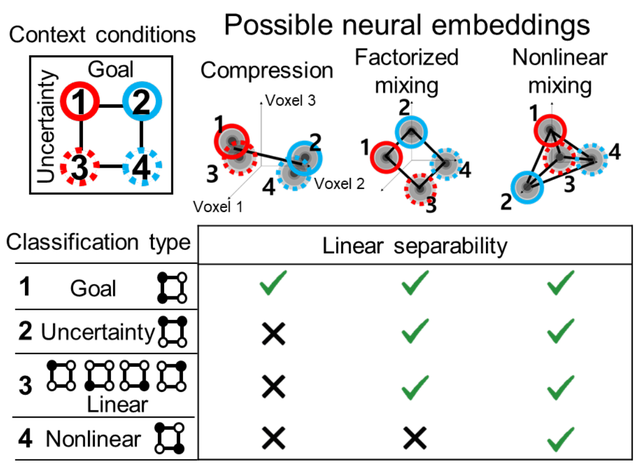
- vlPFC, dlPFC, OFC, and V1: high decoding accruacy for the goals / states
  → LPFC: abstract goal information, V1: sensory state information

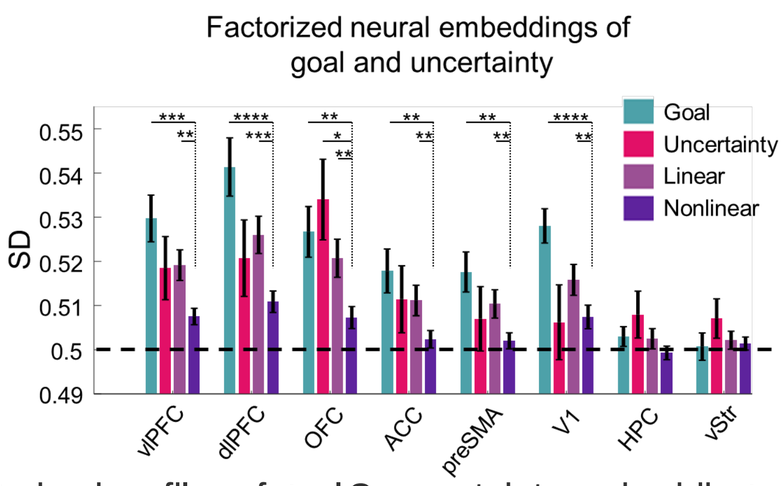


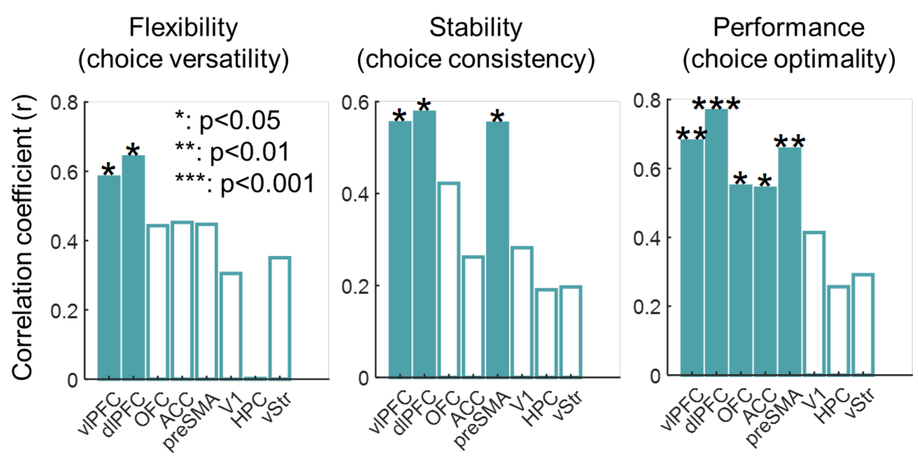
- vlPFC, dlPFC, & OFC: significant uncertainty acc. in the specific goal condition
  → LPFC & OFC: uncertainty-sensitive only during goal-directed behavior

3 Factorized embedding of goal and uncertainty in the LPFC for goal-directed learning

Shattering dimensionality [2, 3] := *average test accuracy of linear dichotomies*



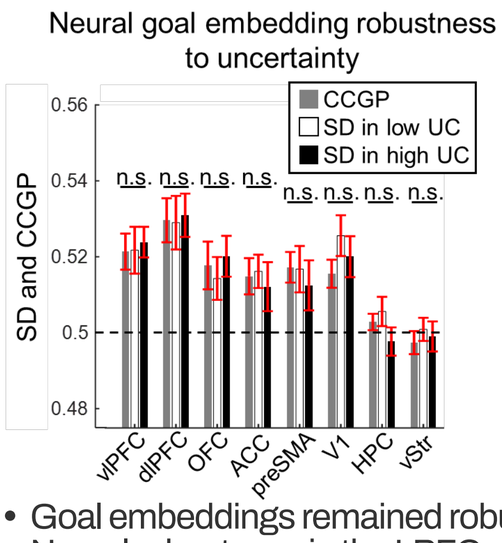


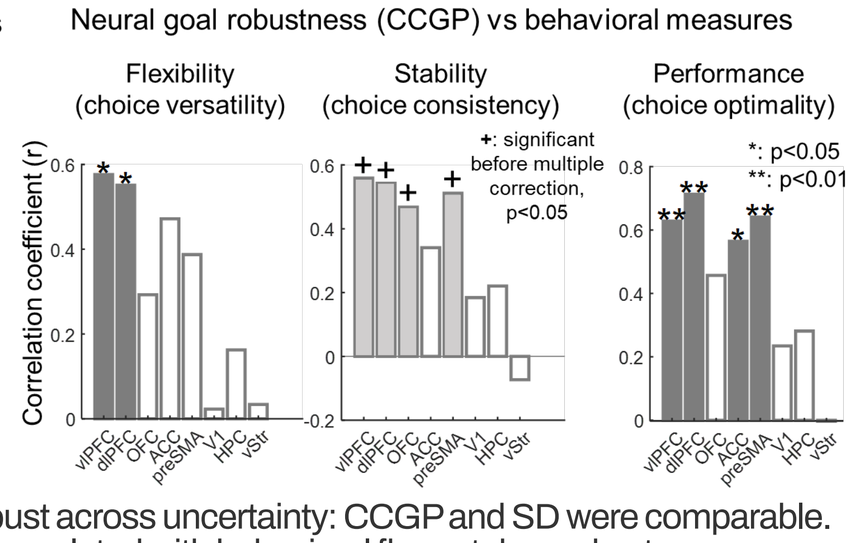


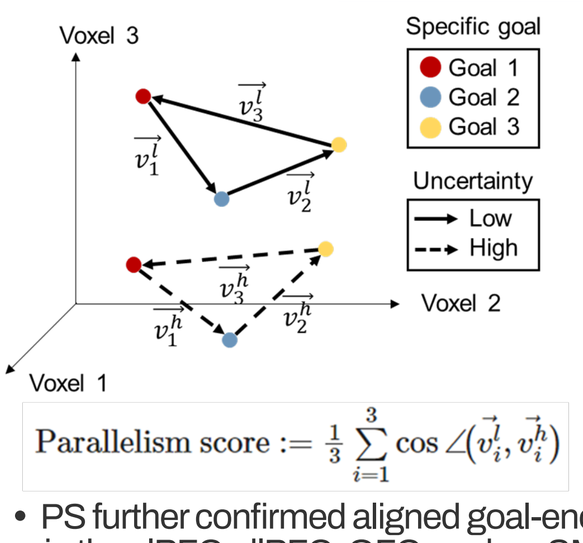
- The vlPFC, dlPFC, and OFC showed factorized profiles of goal & uncertainty embeddings.
- The ACC, preSMA, and V1 exhibited compression, encoding goals alone.
- Neural goal separability in the LPFC predicted individual flexibility, stability, and optimality.

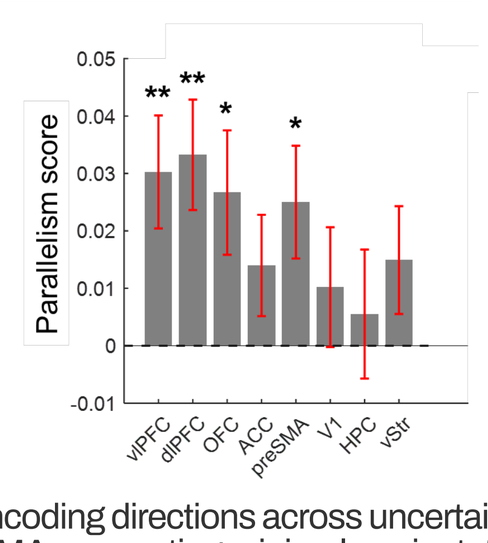
4 Neurally stable goal embedding in LPFC guides stably flexible learning

Cross-condition generalization performance (CCGP) [3] := *average test accuracy of linear classfiers trained in one UC condition and tested in the other condition*









- Goal embeddings remained robust across uncertainty: CCGP and SD were comparable.
- Neural robustness in the LPFC correlated with behavioral flex., stab., and opt.
- PS further confirmed aligned goal-encoding directions across uncertainty in the vlPFC, dlPFC, OFC, and preSMA, supporting minimal reorientation.

References

[1] Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. Neuron, 81(3), 687–699. https://doi.org/10.1016/j.neuron.2013.11.028

[2] Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature, 497(7451), Article 7451. https://doi.org/10.1038/nature12160

[3] Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. Cell, 183(4), 954-967.e21. https://doi.org/10.1016/j.cell.2020.09.031

Acknowledgement

This research was supported by the National Research Foundation of Korea(NRF) funded by the Korean government (MSIT) (No. RS-2024-00439903, RS-2024-00341805, RS-2022-NR068758), and was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00233251, System3 reinforcement learning with high-level brain functions; RS-2019-II190075 Artificial Intelligence Graduate School Program(KAIST); No.2021-0-02068, Artificial Intelligence Innovation Hub). This research was also supported by the MSIT(Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program(RS-2024-00436680) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). This project is supported by Microsoft Research Asia.