# Prefrontal Representations During Learning Reflect Probabilistic Computations Across Domains

**Fahd Yazin (fahd7yazin@gmail.com)**
University of Edinburgh, UK

**Gargi Majumdar (gargi.majumdar@uni-hamburg.de)**
University of Hamburg, Germany

**Neil Bramley (neil.bramley@ed.ac.uk)**
University of Edinburgh, UK

**Paul Hoffman (p.hoffman@ed.ac.uk)**
University of Edinburgh, UK

## Abstract

**The prefrontal cortex (PFC) is thought to represent abstract forms of cognitive maps or internal models during tasks. These representations could be specialized structures suited for distinct domains of experience (e.g., people vs places). Alternatively, they could represent domain-general processes rather than structure, suited for inference across domains. Here we tested these competing accounts using a learning task where human participants learned probabilistic cognitive maps in an unsupervised manner, across three domains, while performing rule classifications. During spatial, social and sequential learning, we found that the structured 1D map representations are formed in the entorhinal cortex but not in midline PFC. Instead, the PFC performs probabilistic inference, abstracting out the underlying probability distributions. Specifically, the ventromedial PFC computes data likelihood under different models, updating them through experience akin to a Bayesian learner. The anteromedial and dorsomedial PFC represent (angular) directional changes and transition distances respectively, within this abstract probability space. These findings were seen during inference as well on unseen exemplars. These results suggest that the midline PFC might be performing a domain-general computation on learned cognitive maps - probabilistic search.**



**Keywords:** Prefrontal Cortex, Cognitive Maps, Bayesian Inference


## Introduction

Cognitive maps or world models are formed from experience, enabling flexible inference from sparse data(Whittington et al., 2022). While these maps are encoded in the medial temporal lobe for physical domains, the prefrontal cortex (PFC) plays an especially key role in abstract domains, crucial for probabilistic decision-making(Constantinescu et al., 2016).

Recent studies show that during naturalistic experiences, PFC regions specialize in updating different parts of the world model ("domains"): current contextual state (vmPFC), others' beliefs (amPFC), and action transitions (dmPFC)(Yazin et al., 2025). This could reflect neural specialization for representing different domains of experience. Alternatively, it could be reflective of different underlying computations required to model different domains. For e.g., coding beliefs-of-others may need computing a reference frame. This is a process-level account, where the representational structure in PFC is reflective of the abstract processes required to extract models from sensory data, rather than true domain-specificity.

Here we pit these two accounts of representation in the PFC by performing functional neuroimaging, when participants learned distinct domains of a virtual world – spatial, social and sequential.

## Methods & Results

We used Age of Empires II to create virtual worlds that combined naturalistic richness with complete experimental control. In our task, 31 participants applied a learned rule to categorize continuously distributed stimuli ($x_1$ or $x_2$, drawn from two overlapping Gaussians) paired with a discrete feature ($y_1$ or $y_2$; e.g., tent or tower) (Fig 1a). While responding, they also implicitly learned the mapping of y1/y2 to the x dimension. To succeed, they had to encode the 1D structure of x along with the underlying Gaussian distributions. Four exemplars per x–y combination (16 total) were presented over 72 learning trials, across 3 blocks. Each domain had unique domain-specific stimulus associations to be learned. Spatial (magnitude of mines - building identity), Social (mental states - behavioural interaction of two people) and Sequential (woodcutting/fire - transport sequence) had thus distinct stimuli/dynamics (Fig 1b). Importantly, the latent structure across three domains remained the exact same. Overall participants learned the two models (Gaussians) along two orthogonal task dimensions, in each domain.

Normalized Beta estimates for exemplars (pooled over blocks) were obtained from a GLM using least squares-single approach, to model the unsmoothed BOLD time-series. fMRI acquisition was through 3T Siemens Skyra, using a multi-echo sequence at 1.7 TR. We then submitted these neural patterns to a wholebrain searchlight representational similarity analysis (RSA), using a range of theoretical dissimilarity matrices (RDMs of dimension 16x16) each operating at varying levels of abstraction (Fig 1). Distance between patterns were computed using Euclidean distance and Spearman's rank correlation was used to estimate similarity between neural and theoretical matrices.

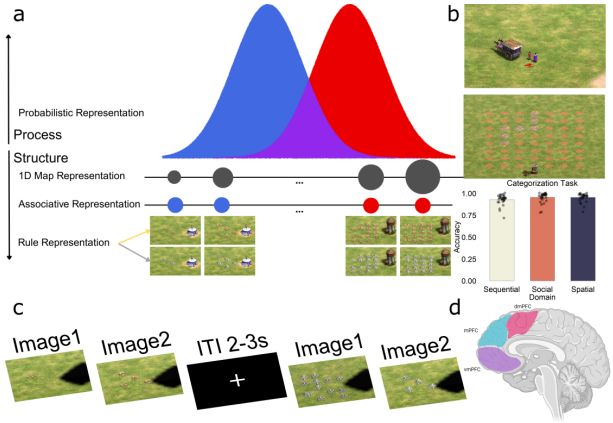


Figure 1: a) Learning task and hierarchy of representational abstractions. b) social, sequential domain examples, and rule learning performance. c) Later Inference task required knowledge of y from two samples of x d) hypothesized prefrontal subregions.

### Structured representation 1: Orthogonal Rule (Fig 2a &d).

This is the task rule, and Euclidean distance between exemplars is computed as

$$|x_i y_1 - x_j y_2| = 0, if\ i = j, else\ 1$$

We found reliable motor cortex patterns consistently across domains for this RDM, suggesting domain-general responses (Fig 2d).

### Structured representation 2: Semantic Association (Fig 2b & e).

Participants represented the binary y1/y2 distinction here.

$$|x_i y_m - x_j y_n| = 0, if\ m = n, else\ 1$$

We found high representational similarity in the visual cortex (Fig 2e) during conceptual domain-specific association, across domains. Patterns were only minimally overlapping.

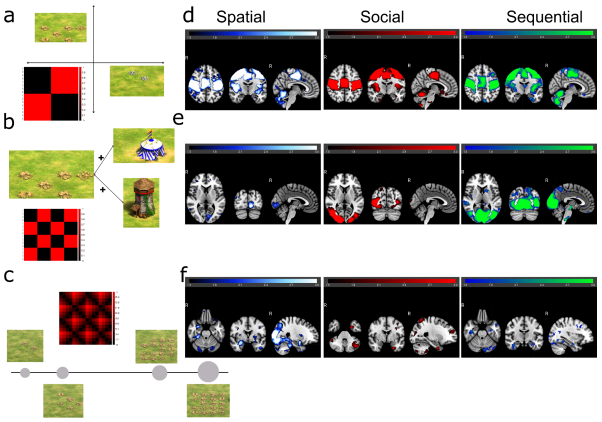


Figure 2: a to c Structured representational hypotheses and RDMs. d to f Whole-brain searchlight RSA performed on spatial (left), social (middle), sequential domain (right) under each RDM.

### Structured representation 3: 1D Cognitive map (Fig 2c & f).

This RDM used the Euclidean distance between exemplars on their continuous dimension (x) to obtain the 1D structure.
In all domains, different segments of Entorhinal cortex showed representational similarity to the general task structure (Fig 2f). Despite this, we did not obtain reliable representations in the PFC for these structured representations during learning. Next, we explore a process account.

### Process representation 1: Likelihood Computation (Fig 3a).

The most abstract task in our study is density estimation, of the two overlapping Gaussians. For this, we hypothesized participants update beliefs rationally like a Bayesian learner. We used conjugate Bayesian update to learn the parameters. At each trial, participants computed the likelihood of the x value under their current set of model parameters μ & σ, for each model (with minimal observational noise).

$$p(x|\theta, m_1) = \frac{1}{\sigma_{m1}\sqrt{2\pi}} e^{\frac{(x-\mu_{m1})^2}{2\sigma^2}}$$

Pooling across all three domains, vmPFC representations correlated with participants' internal representation of the likelihood of observing the data point under the distribution, updating through experience (Fig 3d, top). This suggests participants might be exploring the probability space, searching for models to explain the data.

### Process representation 2 & 3: Distance & Direction Computation in Probability Space (Fig 3b, c).

If participants were indeed searching the probability space, then trial to trial changes to distance and directions should reflect this. Direction was computed by cosine angle between trial vectors(Park et al., 2021) (Fig 3b)

$$\varphi = cos^{-1}\left(\frac{trial_t . trial_{t-1}}{\|trial_t\|\|trial_{t-1}\|}\right)$$

And likewise, distance (Fig 3c) between trials, $A$ by

$$A = |trial_t - trial_{t-1}|$$

We found that amPFC and dmPFC representations specifically were tuned to these changes (3d, middle, bottom). Similar patterns emerged during Inference/test phase, which generalized within the distribution to unseen exemplars (Fig 3e). These results suggest the midline PFC might be computing probabilistic search on the learned maps.

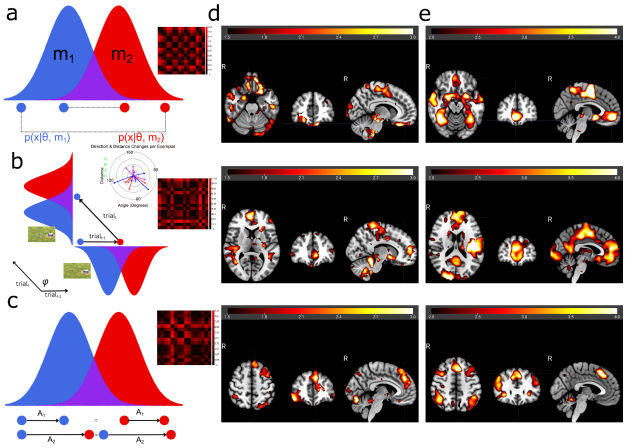


Figure 3: a to c various process representations with example RDM. Polar plot shows a subjects' direction and distance for all exemplars. Learning (d), Inference (e) maps pooled over domains.

Probabilistic search is a foundational computation in machine learning, powered by general-purpose algorithms such as stochastic gradient descent (deep learning), monte carlo tree search (reinforcement learning), and Markov chain monte carlo (Bayesian inference). Our results suggest that the midline PFC representations were similar to a domain-general probabilistic search (Bramley et al., 2023), refining the models through learning.

## Acknowledgments

This project was supported by BBSRC grant (BB/T004444/1).

## References

Bramley, N. R., Zhao, B., Quillien, T., & Lucas, C. G. (n.d.). Local Search and the Evolution of World Models. Topics in Cognitive Science, n/a(n/a). https://doi.org/10.1111/tops.12703

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. Science, 352(6292), 1464–1468. https://doi.org/10.1126/science.aaf0941

Park, S. A., Miller, D. S., & Boorman, E. D. (2021). Inferences on a multidimensional social hierarchy use a grid-like code. Nature Neuroscience, 24(9), 1292–1301. https://doi.org/10.1038/s41593-021-00916-3

Whittington, J. C. R., McCaffary, D., Bakermans, J. J. W., & Behrens, T. E. J. (2022). How to build a cognitive map. Nature Neuroscience, 25(10), 1257–1272. https://doi.org/10.1038/s41593-022-01153-y

Yazin, F., Majumdar, G., Bramley, N., & Hoffman, P. (2025). Fragmentation and Multithreading of Experience in the Default-Mode Network (p. 2024.10.24.620113). bioRxiv. https://doi.org/10.1101/2024.10.24.620113