Controlling PFC dynamics for slow and fast learning

Michał J. Wójcik[1] Jascha Achterberg[1] Joseph Pemberton[2] Rui Ponte Costa[1]
[1]Centre for Neural Circuits and Behaviour, University of Oxford
[2] School of Computer Science and Engineering, University of Washington

Abstract

The prefrontal cortex (PFC) exhibits a remarkable capacity to employ two distinct strategies when engaging in cognitive tasks. Upon encountering a novel task, it leverages high-dimensional representations, well positioned for rapid linear decoding. However, with growing task familiarity, the PFC transitions to employing generalisable low-dimensional neural codes. Through a system-level modelling approach, we propose that these properties emerge naturally in recurrent neural networks (RNNs) that learn on two distinct timescales: (i) on a faster timescale an external controller drives RNN dynamics to generate task-encoding but relatively unstructured, high-dimensional representations, which is then followed by (ii) a slower optimisation of recurrent connections and consequently more structured, low-dimensional representations. We validated these predictions by comparing model representations to neural recordings from the prefrontal cortex of non-human primates that were trained to learn a complex cognitive task from scratch. In summary, our results suggest a learning-dependent control of prefrontal dynamics via a separate brain-region for high-to-low representational switching.

Keywords: PFC; controller, fast and slow learning, neural geometry

Introduction

The geometry of neural representations employed by the brain is a subject of intense debate. In the PFC of performing animals, for example, there is evidence both for high-dimensional and randomly mixed neural selectivity (Rigotti et al., 2013) but at the same time also low-dimensional and highly structured representations (Hirokawa, Vaughan, Masset, Ott, & Kepecs, 2019). It has been suggested that learning plays a key role in determining which of these opposing strategies is used. Specifically, whilst high-dimensional representations may be initially useful for linear separability of novel task features, low-dimensional representations enables better generalisation and reduced metabolic costs.

Here we propose a neural network framework which explains the switch from high to low representations over learning. Specifically, we hypothesise that task learning unfolds in two distinct phases, orchestrated by an external controller. During the initial, fast learning phase, the PFC may operate as a dynamic reservoir supporting rich, high-dimensional representations optimised by external control signals. Crucially, these initial representations are temporary and therefore less constrained by longer-term biological considerations such as metabolic efficiency or robustness to internal noise. In contrast, the subsequent, slower phase is characterised by local synaptic plasticity within the PFC, progressively refining these representations toward lower dimensionality. This refinement explicitly accounts for biologically realistic constraints, ultimately resulting in stable and energy-efficient neural representations suitable for repeated task execution (Stroud et al., 2025; Farrell, Recanatesi, Moore, Lajoie, & Shea-Brown, 2022).

Modelling Framework

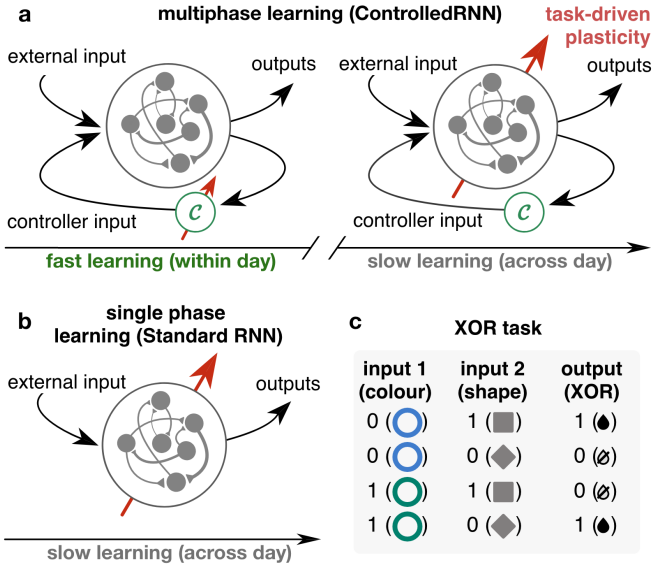

Figure 1: **a**,**b**, Possible learning frameworks. **c**, Structure of the XOR task

To implement multi-phase learning, we consider how representations change within a recurrent neural network whose hidden state $\mathbf{h}$ is driven by external input $\mathbf{x}$ and an external task-optimised signal $\mathbf{c}$. Specifically, we model neural dynamics with

$$\tau\dot{\mathbf{h}} = -\mathbf{h} + \tanh\left(W_{\text{rec}}\mathbf{h} + W_{\text{inp}}\mathbf{x} + W_{\text{c}}\mathbf{c} + \eta\right), \qquad (1)$$

where τ is the membrane time constant ($\tau = 50$ms), W_{rec}, W_{inp}, W_{c} denote the recurrent, input and weights from the controller to the RNN, respectively, and η being a sample from a Gaussian white noise process.

Since $\mathbf{c}$ is used to effectively guide the network to desirable states, it is considered a control signal for the RNN. In this work, we consider $\mathbf{c}$ as the output of a feedforward controller network $\mathcal{C}$ that is bidirectionally connected to the RNN,

$\mathbf{c} = \mathcal{C}(\mathbf{h})$. Crucially, during exposure to a novel task, distinct phases of learning are assumed. In the first phase, which is assumed to be relatively fast (e.g. within one day), controller weights $W_{\mathcal{C}}$ are updated, whilst RNN weights $W_{\text{RNN}} = \{W_{\text{inp}}, W_{\text{rec}}\}$ remain fixed to their randomly initialised values. In the second phase, which is assumed to be slower (e.g. across days), the RNN weights W_{RNN} are now updated. We contrast this controller-based learning strategy with the more standard framework in which an RNN continually learns without an external controller (Fig. 1**a** vs **b**).

Comparison of model activations and nonhuman PFC activity

To evaluate our model's predictions, we drew on a recent experimental study demonstrating learning-related transformations in neural representations (Wojcik et al., 2023). In this study, macaques were trained to perform a context-dependent XOR task, which required non-linear integration of shape and colour cues to predict reward outcomes (Fig. 1**c**). Across multiple sessions, neural recordings from the PFC showed both an enhanced encoding of the XOR, consistent with learning, and a concurrent reduction in dimensionality. To assess whether our framework recapitulates these empirical findings, we trained RNNs on an analogous task. Network weights were updated using gradient descent to minimise task error. Following prior work (Stroud et al., 2025; Whittington, Dorrell, Ganguli, & Behrens, 2023), we incorporated biologically motivated regularisation of both weights and activations during the slow phase of learning, providing a plausible mechanism for dimensionality reduction.

The controller-based framework captures multi-phase learning

The hidden populations both, in the standard RNN and the controlled RNN, rapidly acquired the ability to predict the XOR during the initial "fast phase" of training, mirroring the learning dynamics observed in the PFC within the first day (Fig. 2**a**). To further investigate the underlying neural representations, we examined their geometry and dimensionality. Building on the methods of Bernardi et al. and Wojcik et al., we tested whether the XOR motif was encoded in an abstract format by calculating cross-condition generalisation scores (Fig. 2**b**). Furthermore, we quantified the dimensionality of these neural representations using shattering dimensionality (Rigotti et al.;Fig. 2**c**). In the standard RNN, performance was tightly coupled with an increase in the abstractness of the XOR representation and a concurrent reduction in dimensionality. In contrast, the controller-based model exhibited a distinct decoupling between learning progress and changes in neural geometry and dimensionality. As predicted by reservoir computing, dimensionality increased during the initial "fast phase", where control signals are optimised to manipulate the RNN. This was followed by a significant decrease in dimensionality during the "slow phase" as local weight adjustments within the RNN took place. Notably, this dynamic also influenced the format of the XOR representation - no geometric changes were detected during the "fast" learning phase. Only the controller-based framework reproduced the population dynamics observed in the PFC.

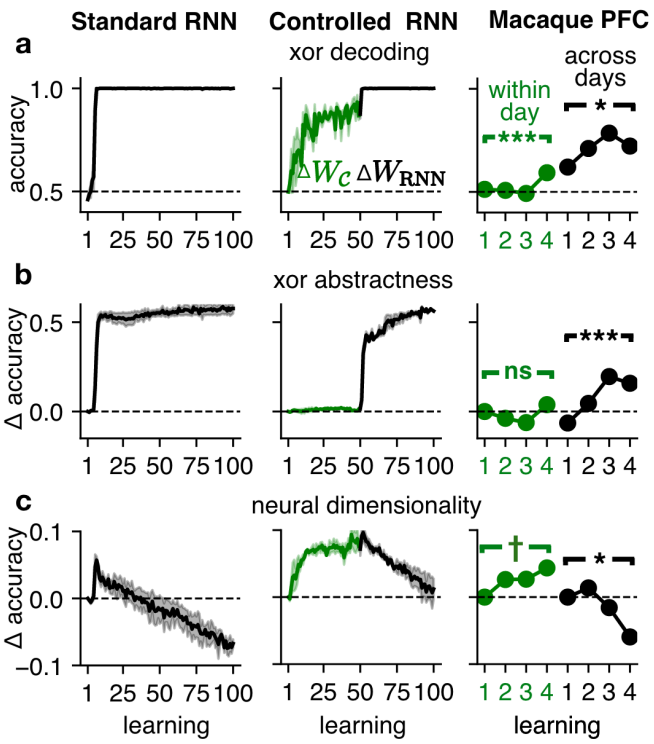

Figure 2: Task representation geometry and dimensionality in artificial and cortical neural networks. **a**, linear XOR decoding; **b**, XOR abstractness as measured by cross-condition generalised decoding; **c**, Neural dimensionality, mean across all possible binary decoders; b, c corrected for initial values.

Discussion

Conceptually, this multi-phase learning approach aligns with theories of memory consolidation, where task-specific knowledge gradually transitions into stable cortical circuits, enhancing generalisation (Sun, Advani, Spruston, Saxe, & Fitzgerald, 2023). Consistent with this, we observed that recurrent weight updates resulted in neural representations that exhibit greater temporal stability and reduced dependence on external control signals (data not shown).

An open question remains regarding how the mammalian brain orchestrates the interplay between fast and slow learning processes. In our model, plasticity arbitrarily shifts from controller-mediated rapid adjustments to slower RNN-based learning, simulating the transition from initial task acquisition to sustained skill refinement analogous to within-day versus across-day learning in animals. The mammalian brain, however, likely employs a more sophisticated architecture, involving distinct yet interacting neural substrates for each learning phase. We speculate that rapid learning predominantly engages supervised or reinforcement-driven mecha-

nisms due to their reliance on immediate error or reward signals. These mechanisms might interact with slower consolidation processes implemented via unsupervised, local learning rules that incrementally refine neural representations over extended timescales (Feulner, Perich, Miller, Clopath, & Gallego, 2025).

References

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*, 1–14.

Farrell, M., Recanatesi, S., Moore, T., Lajoie, G., & Shea-Brown, E. (2022). Gradient-based learning drives robust representations in recurrent neural networks by balancing compression and expansion. *Nature Machine Intelligence*, *4*(6), 564–573.

Feulner, B., Perich, M. G., Miller, L. E., Clopath, C., & Gallego, J. A. (2025). A neural implementation model of feedback-based motor learning. *Nature Communications*, *16*(1), 1805.

Hirokawa, J., Vaughan, A., Masset, P., Ott, T., & Kepecs, A. (2019). Frontal cortex neuron types categorically encode single decision variables. *Nature*, *576*(7787), 446–451.

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), 585–590.

Stroud, J. P., Wojcik, M., Jensen, K. T., Kusunoki, M., Kadohisa, M., Buckley, M. J., … Lengyel, M. (2025). Effects of noise and metabolic cost on cortical task representations. *eLife*, *13*, RP94961. Retrieved from https://doi.org/10.7554/eLife.94961.2 doi: 10.7554/eLife.94961.2

Sun, W., Advani, M., Spruston, N., Saxe, A., & Fitzgerald, J. (2023). Organizing memories for generalization in complementary learning systems. *Nature neuroscience*.

Whittington, J. C. R., Dorrell, W., Ganguli, S., & Behrens, T. (2023). Disentanglement with biological constraints: A theory of functional cell types. In *The eleventh international conference on learning representations*. Retrieved from https://openreview.net/forum?id=9Z$_G$$fhZnGH$

Wojcik, M. J., Stroud, J. P., Wasmuht, D. F., Kusunoku, M., Kadohisa, M., Myers, N. E., … Stokes, M. G. (2023). Learning shapes neural geometry in the prefrontal cortex. *bioRxiv*, 2004–2023.