How the Brain Can Build on Prior Learning: A Model and a Supporting Experiment

Lillian Chang[1], Alexander Farahbod[1], Andrew Kim[1], Christopher Cornell[1], Noah Steinberg[1], Florencia Martinez Addiego[1,2], Yuri Jung[1], Drew Di Donna[1], Tyler Morgan[3], Peter Bandettini[3], Maximilian Riesenhuber[1,4]

[1]Department of Neuroscience, Georgetown University Medical Center; [2]School of Psychology, Georgia Institute of Technology; [3]National Institute of Mental Health; [4]Center for Neuroengineering, Georgetown University

Introduction

Knowledge representation in "deep" AI architectures generally does not lend itself to iteratively build on prior learning: Once a new concept is learned at the top of the hierarchy, there is no mechanism to explicitly integrate this learned concept into the existing hierarchy to permit the system to build on it in future learning.

In contrast, humans can leverage prior knowledge to facilitate the learning of novel tasks, as in the phenomenon known as "fast mapping".

Prefrontal cortex (**PFC**): shape processing. Plays a key role in flexibly learning new tasks[1-4]. However, PFC represents a computational bottleneck; task execution is limited to one task at a time.

Anterior temporal lobe (**ATL**): key area for semantic processing in the human brain. Evidence from fMRI and EEG for category selectivity after 30,000 trials[5].

Question: To avoid the computational bottleneck after initial learning in the PFC and facilitate faster future learning, what kind of model is needed to represent transfer of learned category information?

Hypothesized model:

Unsupervised learning of shape representations in the deep sensory processing hierarchy (dotted arrow; Fig. 1).

Supervised learning of concept judgment circuits in PFC[1-4]. ATL is still only tuned to specific objects.

PFC representations provide re-entrant supervised teaching signals to ATL neurons during *correct* trials (not error-driven), broadening ATL unit selectivity into concept selectivity and thereby achieving transfer task automation[5]. Task circuits transferred to ATL can serve as input to PFC in subsequent task learning."

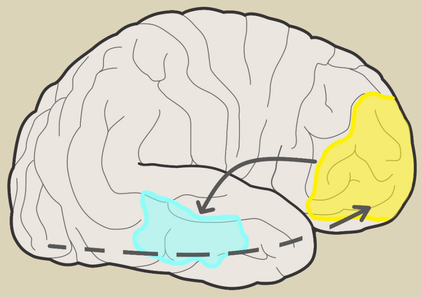


Fig 1. Graphical representation of our model of task transfer. PFC (yellow) initially learns the task, then transfers to ATL (cyan).

Experimental Design: Categorization of Morphed Cars

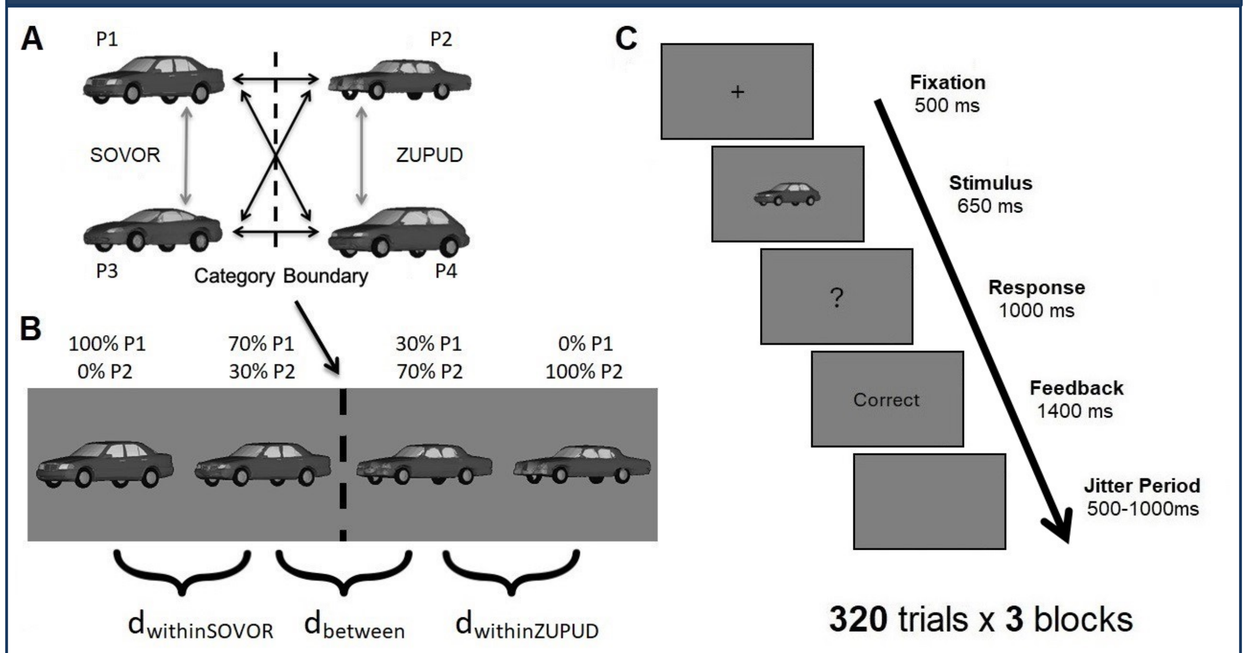
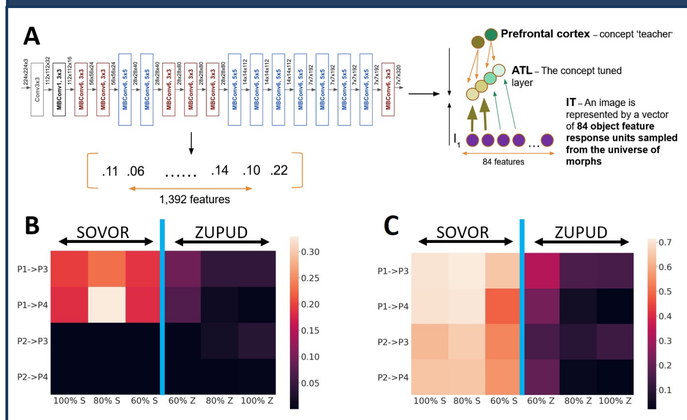


Fig 2. EEG experimental paradigm. **A.** Category designations illustrating the 4 prototype stimuli[2,5] (P1-4). An additional twelve cars were created by morphing between two cars. **B.** Example of the morph process between two prototypes. With 4 prototypes, and 12 morphs = 16 unique stimuli; an optimal number to be learned in a single session that allows us to track transfer learning of category circuits. For our dissimilarity analysis (EEG-RSA; Fig. 6), the crossnobis distance per block required calculating the differences for within-category stimuli ($d_{withinSOVOR}$ & $d_{withinZUPUD}$) and between/across category stimuli ($d_{between}$). **C.** Trial structure. In total, 25 participants completed the task while neural data was collected from 128 scalp electrodes, acquired using a BioSemi EEG system sampled at 2048 Hz.

Computational model: Transfer of learned category tasks from "prefrontal" to "anterior temporal circuits" by virtue of re-entrant signals



- Used EfficientNetB3 – layer 'block6e_se_squeeze' with 1,392 features to extract a dense feature representation of an image (Fig. 3A).
  - PFC units already tuned – categories already learned.
  - Initially, ATL units are only selective for exemplars Fig. 3B), like inferior temporal cortex (IT) neurons.
  - Re-entrant signals from PFC gradually tune ATL units to respond similarly to cars from the same category, but differently to cars from different categories (Fig. 3C).

Fig 3. Computational model simulations. **A.** A convolutional neural net (EfficientNetB3) was used to create the initial, baseline, feature vector. Simulations show how re-entrant signals from category-selective PFC units can broaden initial object selectivity of ATL units (**B**, with the model unit initially only showing responses to morphs involving a single prototype) into concept selectivity (**C**, with strong responses to all "SOVOR" members).

Behavioral and EEG Analyses Support Transfer of Learned Concepts

Participants Demonstrate Fast Learning and Asymptotic Behavior

Performance on across-boundary morphs saturates within the second block (1st vs. 2nd half).

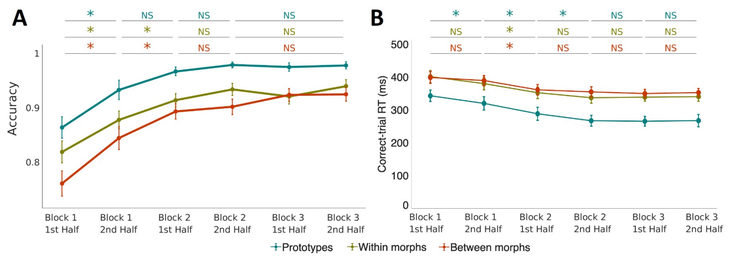


Fig 4. Behavioral performance from N=25 participants, separated by stimuli type, measured by **A.** accuracy (two-tailed paired t-tests) and **B.** reaction time (one-tailed paired t-tests). NS: p≥0.05, *p<0.05.

Sensors Corresponding to Regions of Interest

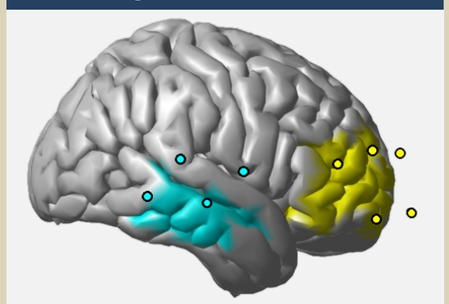
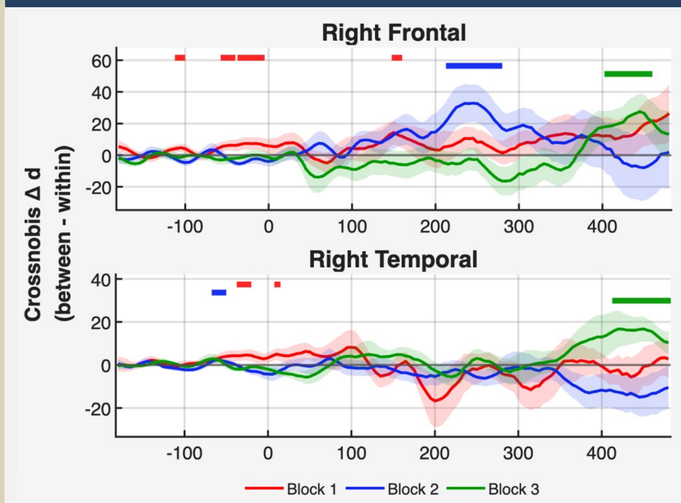


Fig 5. Right hemisphere anatomical target ROI were defined then used in source space beamforming analyses. Scalp EEG sensors that contributed the most activation to each ROI were selected for sensor space analyses. Yellow=frontal, corresponding to PFC. Cyan=temporal, corresponding to ATL.

Representational Similarity Analysis (EEG-RSA)



Look at differences between the average dissimilarity distance, calculated with a cross-validated squared Mahalanobis (crossnobis), for between- vs. within-category morphs.

- Distance=0: no special status in the boundary.
- Difference>0: Greater dissimilarity for between-boundary morphs → stimuli have been learned into their respective categories.

Results

Increase in distance separability reflects category learning in each region.

- Frontal first in block 2, then block 3 for temporal.

Fig 6. Dissimilarity analysis of neural responses to across-boundary morph stimuli in each ROI. Data was epoched around stimulus onset time. Sliding window=40ms. Two-tailed one-sample t-tests (p<0.05).

Granger Causality Analysis Supports Frontal→Temporal

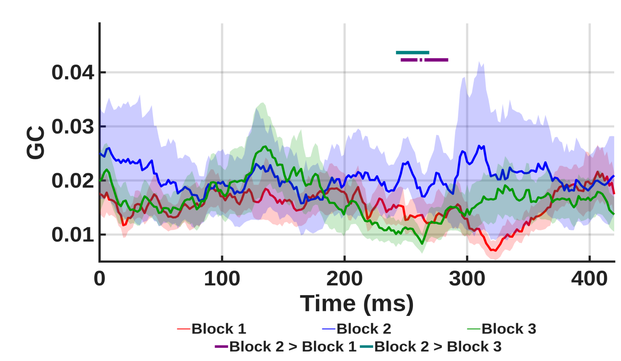


Fig 7. Sensor space Granger Causality (GC) analysis of across-boundary morph trials. Data was downsampled to 500 Hz to calculate frontal to temporal GC modulations in the theta (4-7 Hz) band for right hemisphere frontal and temporal sensors. Sliding window=40ms. Model order=15. One-tailed paired t-tests (p<0.05).

Prediction: 1) GC from frontal→temporal initially low as participants learn the task. 2) After frontal circuits learn the task, these task circuits provide re-entrant teaching signals to offload the task to temporal cortex—increase in GC from frontal→temporal channels. 3) Frontal modulations of temporal cortex decrease after task transfer, leading to a decrease in frontal→ temporal GC.

Results: Significant increase of frontal→temporal in block 2 around 250-300ms, then significant decrease in block 3 as predicted; coinciding with the timing seen in the EEG-RSA analyses.

Summary

In this study, we proposed a mechanism to build on prior learning and incrementally learn more complex tasks, with implications for automating task execution.

By conducting analyses that examined how neural selectivity in frontal areas modulates temporal units and calculating differences in dissimilarity over time, our computational model and EEG results support the proposed conceptual model of how the brain offloads task circuits from frontal to temporal regions.

Notably, the model is biologically plausible and does not require the propagation of error-related teaching signals, evidence for which has remained elusive.

Future directions:

Examine the neural dynamics underlying our model at a finer spatial resolution using layer-fMRI.

Acknowledgements

Research funded by the Army Research Office (W911NF-24-1-0097).

References

1. Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, *23*(12), 5235-5246. https://doi.org/10.1523/JNEUROSCI.23-12-05235.2003.
2. Jiang, X., Bradley, E., Rini, R. A., Zeffiro, T., VanMeter, J., & Riesenhuber, M. (2007). Categorization training results in shape-and category-selective human neural plasticity. *Neuron*, *53*(6), 891-903. https://doi.org/10.1016/j.neuron.2007.02.015.
3. Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, *3*(11), 1199-1204. https://doi.org/10.1038/81479.
4. Scholl, C. A., Jiang, X., Martin, J. G., & Riesenhuber, M. (2014). Time course of shape and category selectivity revealed by EEG rapid adaptation. *Journal of Cognitive Neuroscience*, *26*(2), 408–421. https://doi.org/10.1162/jocn_a_00477.
5. Cox, P. H., Scholl, C. A., Laws, M. L., Jaimes, N. E., Jiang, X., & Riesenhuber, M. (2026). Extensive Experience Remodels Neural Task Circuitry to Escape the Frontal Bottleneck and Increase Automaticity of Categorization. *Journal of Cognitive Neuroscience*, 1-23. https://doi.org/10.1162/JOCN.a.2618.