From System 2 to System 1: Neural mechanisms of deep concept learning in the brain

Lillian Chang[1], Jonathan Yu[1], Alexander Farahbod[1], Basith Faizal[1], Andrew Kim[1], Noah Steinberg[1], Ran Yuan[1], Jessica Lin[1], Chris Cornell[1], Florencia Martinez Addiego[1], Laurentius Huber[2], Tyler Morgan[2], Peter Bandettini[2], Maximilian Riesenhuber[1]

[1]Georgetown University - Department of Neuroscience, Center for Neuroengineering; [2]National Institute of Mental Health

Introduction

- Object categorization is a crucial cognitive operation and fundamental to mapping sensory stimuli to semantic knowledge.
- Prior work has shown a key role in the prefrontal cortex (PFC) in the flexible learning of new tasks, including learning new object categories[1-4]. Yet, PFC processing is subject to cognitive control and attentional limitations, limiting its ability to automatically map arbitrary sensor stimuli to meaning.
- In contrast, influential theories of semantic cognition have proposed that the anterior temporal lobe (ATL) may act as a "semantic hub"[5] on top of the ventral stream's simple to complex" processing hierarchy, linking shape to meaning[6-8]. Yet it is less clear how semantic representations are learned in the ATL.
- Hypothesis: Slower, more effortful category learning first occurs in the PFC with its ability to flexibly learn new task modules (akin to Kahnemann's[9] "System 2"). Once category circuits have been learned in the PFC, top-down, re-entrant signals from PFC to ATL mediate transfer learning of these category circuits to ATL (System 1) to allow fast, automatic categorization, freeing up frontal cortex.
- Critically, to track the hypothesized transfer from PFC to ATL, we perform EEG recordings *while* participants are learning.

Experimental Design: Categorization of Morphed Cars

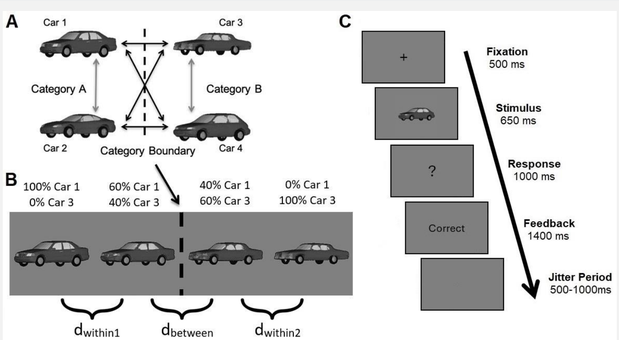


Fig 1. EEG experimental paradigm. **A.** Category designations illustrating the four prototype stimuli[1,3]. An additional twelve cars were created by morphing between two cars. An example of the morph process between two prototypes is shown in **B.**. With 4 prototypes, and 12 morphs, this results in 16 unique stimuli; an optimal number to be learned in a single session that allows us to track transfer learning of category circuits. **C.** Trial structure. Each run consisted of 320 trials. In total, 14 participants completed 3 blocks of the task while neural data was collected from 128 scalp electrodes. Data were acquired using a BioSemi EEG system.

Results

Participants Demonstrate Fast Learning of Novel Categories

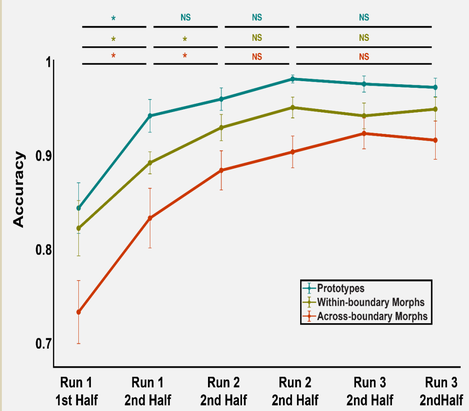


Fig 2. Behavioral performance from N=14 participants, separated by stimulus type. Crucially, performance saturates within the second block. Two-tailed paired t-tests. NS: p≥0.05, *p<0.05.

Sensors Corresponding to Regions of Interest

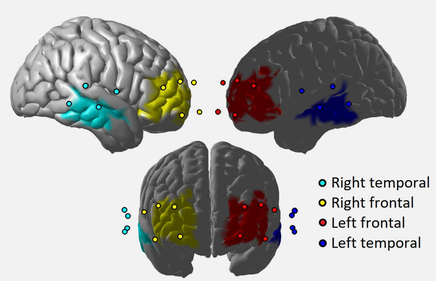
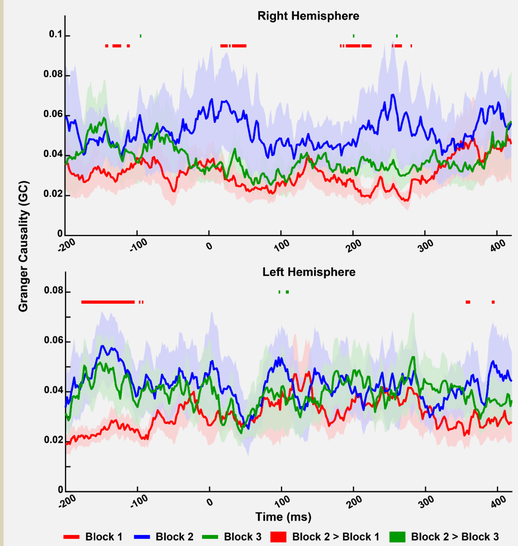


Fig 3. Anatomical target ROI were defined then used in source space beamforming analyses. Scalp EEG sensors that contributed the most activation to each ROI were selected for sensor space analyses.

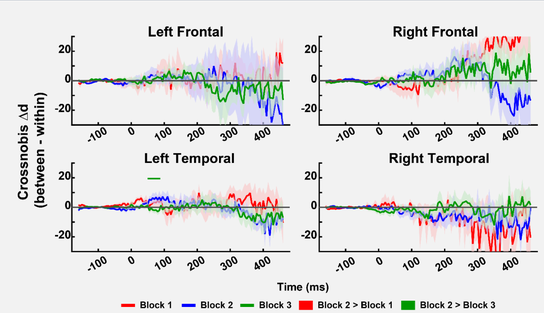
Granger Causality Analysis Demonstrates Modulation from Frontal→Temporal



- Prediction: GC is initially low as participants learn the task. After frontal (System 2) learns the task, it offloads to temporal (System 1) reflected in an increase in GC in the beta band, which has been hypothesized to reflect top-down influences[9]. After frontal has transferred the task to temporal, frontal modulations decrease and beta GC will decrease again.
- We find significant modulation of frontal→ temporal beta from 0-300ms in the right but not left hemisphere, suggesting a key role of the right hemisphere (as shown in Jiang et al., (2007)).

Fig 4. Sensor space Granger Causality (GC) analysis of across-boundary morph trials. GC was calculated for frontal→temporal in the beta (13-20 Hz) band for the **A.** right hemisphere and **B.** left hemisphere frontal and temporal sensors. Model order=15.

Representational Similarity Analysis Shows Differential Learning Patterns for Frontal and Temporal



- Prediction: Increases in category selectivity should show as an increase in the distance comparing across category vs. within category stimulus pairs.
- Distance for between minus within increases first in right frontal.
- Then, dissimilarity increases over blocks only for right temporal, and no change for the left hemisphere.

Fig 5. Dissimilarity analysis of neural responses to across-boundary morph stimuli in each ROI. A crossnobis distance calculation was used to differentiate dissimilarity from noise.

Summary

- We conducted a high-density EEG study to test the hypothesis that the PFC mediates the initial flexible learning of new categories, followed by transfer to temporal circuits. This provides a neural substrate for the learning of new concepts and their automatic execution.
- By conducting analyses that examined how frontal cortex modulates temporal cortex and calculating differences in dissimilarity over time, our results provide preliminary support for the System 2→System 1 transfer hypothesis, with the frontal cortex (System 2) involved in initial category learning, followed by transfer to temporal cortex (System 1).
- Computational simulations show that such a learning scheme allows incremental learning, i.e., the sequential learning of new tasks and their storage in ATL.

Acknowledgements

Study funded by the Army Research Office (AWD-7776145).

References

1. 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.
2. Jiang, X., Chevillet, M. A., Rauschecker, J. P., & Riesenhuber, M. (2018). Training humans to categorize monkey calls: auditory feature-and category-selective neural tuning changes. *Neuron, 98*(2), 405-416.
3. 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.
4. Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science, 291*(5502), 312-316.
5. Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature reviews neuroscience, 18*(1), 42-55.
6. Damera, S. R., Martin, J. G., Scholl, C., Kim, J. S., Glezer, L., Malone, P. S., & Riesenhuber, M. (2020). From shape to meaning: Evidence for multiple fast feedforward hierarchies of concept processing in the human brain. *Neuroimage, 221*, 117148.
7. Jefferies, E. (2013). The neural basis of semantic cognition: converging evidence from neuropsychology, neuroimaging and TMS. *Cortex, 49*(3), 611-625.
8. Chen, Y., Shimotake, A., Matsumoto, R., Kunieda, T., Kikuchi, T., Miyamoto, S., ... & Ralph, M. L. (2016). The 'when'and 'where'of semantic coding in the anterior temporal lobe: Temporal representational similarity analysis of electrocorticogram data. *Cortex, 79*, 1-13.
9. Kahneman, D. (2011). Fast and slow thinking. Allen Lane and *Penguin Books, New York*.
10. Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J. M., Oostenveld, R., Dowdall, J. R., ... & Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron, 85*(2), 390-401.