Neural generalization principles of working memory in humans and recurrent neural networks

Dongping Shi[1,2], Luchengchen Shu[1,2], Qing Yu[1,*]

[1]Institute of Neuroscience, State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China
[2]University of Chinese Academy of Sciences, China
*Corresponding author

Abstract

A fundamental endeavor in cognitive neuroscience is to understand how information can be rapidly abstracted through shared perceptual or structural knowledge to facilitate efficiency and learning. Working memory (WM) provides a flexible mental workspace for these computations, yet how generalization is realized within WM remains largely unexplored. Here, using functional MRI (fMRI) and recurrent neural network (RNN) modeling, we investigated how stimulus and rule information generalize within WM. Across two experiments, participants performed two WM tasks with shared stimulus structure but distinct stimulus sets (location and object), either without (Experiment 1) or with (Experiment 2) explicit mapping. In each task, they flexibly switched between maintenance and manipulation of stimulus information following task rules. Leveraging multivariate decoding and state space analyses, we revealed separate neural substrates in the generalization of stimulus and rule information in WM: the posterior parietal cortex represented mnemonic information across stimulus domains, with enhanced generalization of mnemonic information during memory manipulation compared to maintenance. In contrast, frontal subregions encoded abstract rules that were generalizable across tasks. RNN simulations replicated the key generalization patterns. Together, our findings reveal the neural generalization principles of WM that enable flexible maintenance and manipulation of information for goal-directed behavior.

Keywords: working memory; neural generalization; parietal cortex; prefrontal cortex; fMRI; recurrent neural network

Introduction

In daily life, we often perceive and memorize the external world not in its original form but rather in a modified manner. Information can be abstracted, particularly through some shared perceptual or structural knowledge, to facilitate efficiency and learning (Behrens et al., 2018; Summerfield et al., 2020). The process through which shared information is extracted is referred to as generalization. While the question of how generalization can be implemented in the brain has received growing attention in the recent years, the mechanism underlying rapid generalization, which occurs within a short time window on-task, remains largely unexplored. Working memory (WM), the ability to flexibly maintain and manipulate information to guide behavior (D'Esposito & Postle, 2015), provides a flexible mental workspace for such computations. In this study, we investigated generalization within WM by focusing on two key components of WM: stimulus and rule information. Specifically, stimulus information reflects the mnemonic content held in WM, whereas rule information represents the abstract task-related constraints that acts on the specific content to guide behavior.

Methods

In Experiment 1, human participants (N = 23) completed two WM tasks involving two distinct circular stimulus spaces: location and object (Li et al., 2020). In the location task (Figure 1A), participants mentally maintained or manipulated spatial locations by a cued angle (rotating 0, ±60, ±120, ±180 degrees)(Shi & Yu, 2024). In the object task (Figure 1B), participants first acquired the structure of a circular object space by learning the transitional relations between objects drawn from the space, and during the main task, mentally maintained or manipulated objects according to symbolic cues indicating the stepwise distances to be updated (0, ±1, ±2, ±3 steps). In other words, the two tasks shared both a similar circular stimulus structure and a similar rule structure but with distinct stimulus sets. Experiment 2 followed the same procedure as Experiment 1, except that participants (N = 23) formed a fixed one-to-one mapping between location and object stimuli through behavioral training. This design aimed to form an explicit mapping between stimulus and rule in the two tasks to enable direct comparison between the neural codes.

To examine the neural codes for generalization at the mechanistic level, we simulated RNNs (n = 20) either with or without explicit mapping (Figure 1C) to perform both WM tasks (Masse et al., 2019).

Representational similarity analysis (RSA) was used to decode stimulus representation in each task separately, and subspace decomposition and cross-decoding analysis were used to determine whether there was a generalized code for each condition.

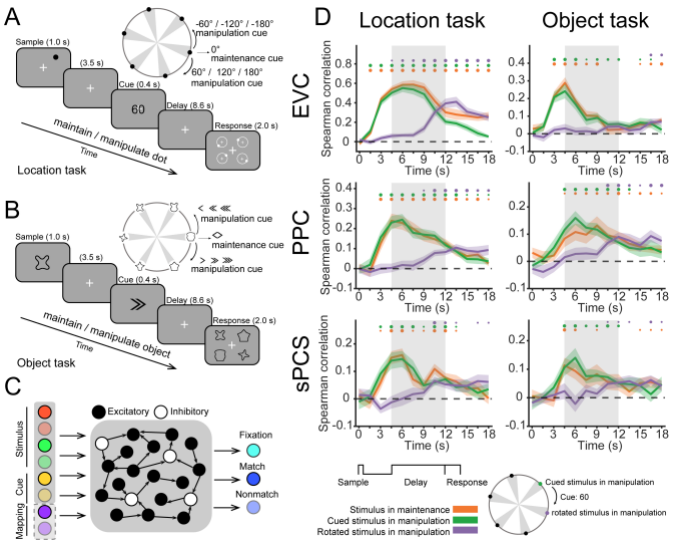

Figure 1. (A–B) Experimental procedure of the location task (A) and object task (B). (C) Architecture of the RNN. (D) Stimulus decoding results in EVC, PPC, and sPCS for Experiment 1. The gray area indicates the delay period.

Lastly, we applied a similar analytic approach to rule information in WM. Intriguingly, we found that rule information was more generalizable in sPCS instead of PPC (Figure 2D). These results suggest separate cortical substrates for stimulus and rule generalization during WM.

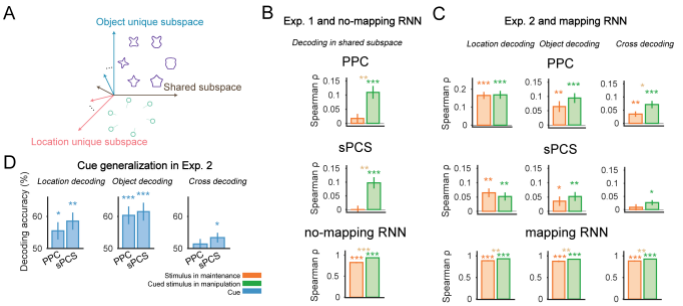

Figure 2. (A) Schematics of subspace decomposition. (B) Stimulus decoding results for location and object tasks in the shared subspace for Experiment 1 and no-mapping RNN. (C) Stimulus decoding and cross-decoding results for Experiment 2 and mapping RNN. (D) Rule decoding and cross-decoding results in Experiment 2 using SVM.

Results

We first identified brain regions that were engaged in both tasks. In Experiment 1, we observed that among the WM-related areas, the posterior parietal cortex (PPC) showed the most persistent stimulus representation for both cued and rotated information across tasks. The superior precentral sulcus (sPCS) in the frontal cortex demonstrated a weaker pattern, whereas the early visual cortex (EVC) failed to exhibit robust coding of the rotated information in the object task (Figure 1D). These results indicate that PPC serves as a domain-general brain region for WM. We then used PCA to decompose PPC activity into three subspaces (Figure 2A): one shared by both tasks and two unique to each task. We found that the neural representation of the cued stimulus in the manipulation condition was stronger in the shared subspace compared to the maintenance condition (Figure 2B), suggesting that active manipulation of information enhanced the generalization of mnemonic information in the PPC.

In Experiment 2, we used a cross-decoding approach and again revealed higher cross-decoding performance in the PPC for the cued stimulus during manipulation, reaffirming that mnemonic generalization is facilitated through manipulation, even when the two stimulus spaces were explicitly linked (Figure 2C). A weaker pattern was observed in the sPCS (Figure 2B-C). Simulations using RNNs replicated the human neural patterns, with enhanced neural representation during manipulation in the shared subspace (Figure 2B).

Discussion and Conclusion

Through two fMRI experiments and RNN simulations, we provide converging evidence for two generalization principles of WM. First, the neural locus for stimulus and rule generalization in WM is spatially separable: stimulus generalization was found to be most robust in the PPC, a brain region that has been implicated in WM (Xu, 2017) and structural learning (Summerfield et al., 2020). In contrast, rule generalization primarily engaged the frontal cortex, a brain region that is critical for WM and cognitive control (Miller & Cohen, 2001; Zhang & Yu, 2024). These results indicate that the generalization of information in WM is highly dependent on the specialized neural modules that actively process the relevant information.

Second, manipulation functions of WM, which require the active control and on-line processing of the maintained information (Shao et al., 2024), can facilitate neural generalization across tasks, regardless of whether the two task spaces were explicitly linked or not. This is likely because manipulation can facilitate the exploration of the structural relationships between stimuli, leading to better extraction of task regularities and ultimately more efficient task performance through generalized representations.

In summary, we reveal that rapid, flexible generalization in WM is realized via a distributed WM network, with different cortical regions specialized in distinct aspects of WM information processing.

References

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron, 100*(2), 490-509. doi:10.1016/j.neuron.2018.10.002

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annu Rev Psychol, 66*, 115-142. doi:10.1146/annurev-psych-010814-015031

Li, A. Y., Liang, J. C., Lee, A. C. H., & Barense, M. D. (2020). The validated circular shape space: Quantifying the visual similarity of shape. *J Exp Psychol Gen, 149*(5), 949-966. doi:10.1037/xge0000693

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X. J., & Freedman, D. J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nat Neurosci, 22*(7), 1159-1167. doi:10.1038/s41593-019-0414-3

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu Rev Neurosci, 24*, 167-202. doi:10.1146/annurev.neuro.24.1.167

Shao, Z., Zhang, M., & Yu, Q. (2024). Stimulus representation in human frontal cortex supports flexible control in working memory. *Elife, 13*, RP100287. doi:10.7554/eLife.100287.3

Shi, D., & Yu, Q. (2024). Distinct neural signatures underlying information maintenance and manipulation in working memory. *Cereb Cortex, 34*(3). doi:10.1093/cercor/bhae063

Summerfield, C., Luyckx, F., & Sheahan, H. (2020). Structure learning and the posterior parietal cortex. *Prog Neurobiol, 184*, 101717. doi:10.1016/j.pneurobio.2019.101717

Xu, Y. (2017). Reevaluating the Sensory Account of Visual Working Memory Storage. *Trends Cogn Sci, 21*(10), 794-815. doi:10.1016/j.tics.2017.06.013

Zhang, M., & Yu, Q. (2024). The representation of abstract goals in working memory is supported by task-congruent neural geometry. *PLoS Biol, 22*(12), e3002461. doi:10.1371/journal.pbio.3002461