Decoding Signatures of Meta-Learning Abstract Structure

Doris Yu (my2689@nyu.edu)
Psychology Department, New York University
New York, NY, USA

Sreejan Kumar (sreejank@princeton.edu)
Princeton Neuroscience Institute, Princeton University
Princeton, NJ, USA

Marcelo G. Mattar (marcelo.mattar@nyu.edu)
Department of Psychology, New York University
New York, NY, USA

Abstract

Humans excel at learning abstract structure from limited data and applying it to novel situations—a capacity often attributed to meta-learning. While behavioral evidence supports this ability, the neural mechanisms by which abstract concepts are acquired and refined during learning remain unclear. In this study, we use magnetoencephalography (MEG) to examine how the brain dynamically constructs abstractions while learning a set of tasks generated with a compositional grammar. Through MEG decoding, our results show evidence of learning the grammar structure across multiple timescales, both within and across different trials. These findings provide neural evidence for meta-learning in humans, showing that abstract representations emerge during learning.

Keywords: Meta Learning, Magnetoencephalography (MEG)

Introduction

The brain has an impressive capability of learning from a sparse amount of data and generalizing our experiences to a wide variety of concepts . Psychologists have long suggested that this ability relies on forming "generalizing abstractions" by identifying patterns across experiences (Hull, 1920). Meta-learning formalizes this idea, showing how abstract concepts learned across tasks can accelerate future learning (Wang, 2021). However, most studies on task-specific abstraction focus on behavior alone (Barack, Bakkour, Shohamy, & Salzman, 2023; Braun, Mehring, & Wolpert, 2010; Gershman, Blei, & Niv, 2010), while neuroscientific studies often record brain activity only after participants have already mastered the task (Schwartenbeck et al., 2023; Bernardi et al., 2020; Tafazoli et al., 2024). This leaves a gap in understanding how new abstract concepts form during learning at the neural level.

In this study, we study how humans naturally meta-learn new abstract concepts when engaged in tasks that have common structure. Specifically, we propose that humans implicitly improve their learning efficiency by inferring the underlying generative model of their experiences. To test this, we employ magnetoencephalography (MEG) to record brain activity while participants engage in a family of tasks that are generated through an abstract compositional grammar. Our results show neural evidence of humans learning the underlying generative grammar to improve their performance over the course of the experiment.

Method

We used a compositional generative grammar to create structured two-dimensional binary grids ("boards") with red and blue tiles. This grammar allows for the generation of stimuli that vary in complexity and structure, including tree, loop, and line-like patterns (Fig. 1ABC). Each board was generated through recursive production rules, resulting in a distribution of samples that varied in both structure and size (see examples in Fig.1C). Following Kumar et al.(2021), we designed a task in which participants viewed a "covered" version of each board (e.g., all tiles initially grey) and revealed tiles one at a time by clicking. Their goal was to uncover all red tiles while minimizing the number of blue tiles revealed.

We first validated learning in a behavioral pilot with 25 participants in each condition (compositional and metamer), each completing 24 tile-revealing trials. Compositional boards were generated from a structured grammar designed to reflect abstract concepts (tree, loop, and line). Following prior work (Kumar et al., 2021), the metamer boards matched the compositional boards in low-level statistics but lacked any apparent high-level structure (Fig.1C). Consistent with previous findings (Kumar, Dasgupta, Cohen, Daw, & Griffiths, 2021), participants performed significantly better in the compositional condition than in the metamer condition (Fig.1D), suggesting stronger learning when abstract structure was present (Fig. 1E). However, these results alone cannot determine whether participants learned the true underlying grammar or formed other sufficient abstractions to support performance.

To study the neural basis of abstraction learning, we recruited participants to complete the same task while undergoing MEG scanning. Each trial began with a random red tile revealed. Participants used a MEG-compatible keypad to navigate the board with directional keys and pressed a separate key to reveal tiles. They earned one point for each new red tile uncovered and lost one point for revealing a white tile. Trials lasted up to 15 seconds, with a visible countdown bar. If completed early, the board remained visible for the rest of the trial; otherwise, the full board was shown for 1 second at the end. Each participant completed 12 blocks of 10 trials and was compensated based on performance.

MEG Setup In this study, we used a whole-head KIT MEG system equipped with 157 axial gradiometers, sampled at 1 kHz from n = 5 subjects.

Decoding Analysis To assess whether participants encoded the compositional structure of the generative rule during learning, we decoded two features from the MEG data: (1) the grammar rule used to generate each board (chain, tree, or loop; Fig. 1B), and (2) the board's generative size (small, medium, or large), based on the number of rule applications in the generative process. For example, two loops or a chain with three tiles is considered small, while four loops or a tree with four expanded branches is considered large. Decoding focused on a 1-second window surrounding each correct reveal event (i.e., when a red tile was uncovered). At each timepoint within this window, we trained a logistic regression classifier using a Leave-One-Trial-Out cross-validation approach. To prevent overfitting, we apply $L2$ regularization, with the regularization parameter selected via a nested cross-validation loop within the training set.

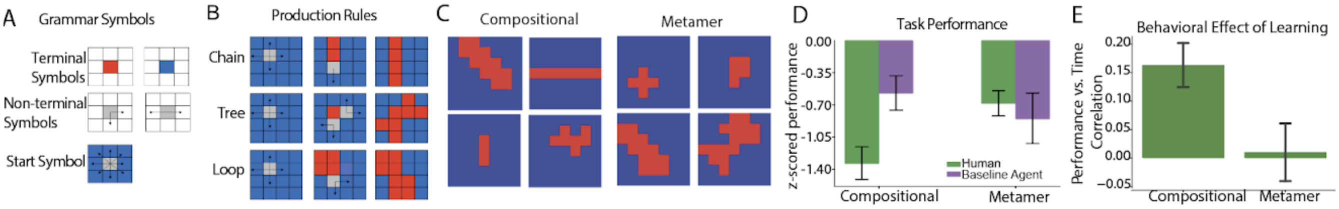

Figure 1: A Novel Compositional Grammar for Board Stimuli. (A) Grammar symbols and (B) production rules. There are three different rules to produce progressively more complex grid structures, and each rule produces a qualitatively different grid structure. These structures can vary in size based on how many times that grammar rule is applied. At the start state, one of the three grammar rules is randomly chosen and applied. This grammar rule is continuously applied until random termination. (C) Example compositional grammar samples vs metamer samples using the same process to generate metamers. (D). Performance of humans and artificial agent on the compositional grammar boards vs metamer boards. The performance metric is based on the number of white tiles revealed, so lower is better. Error bars are 95% confidence intervals over different participants (human), and baseline model runs. (E) Humans learn significantly over the course of the experiment (indicated by correlation between trial number and performance), more so for compositional tasks than metamer tasks (p=0.0006).

Results and Discussion

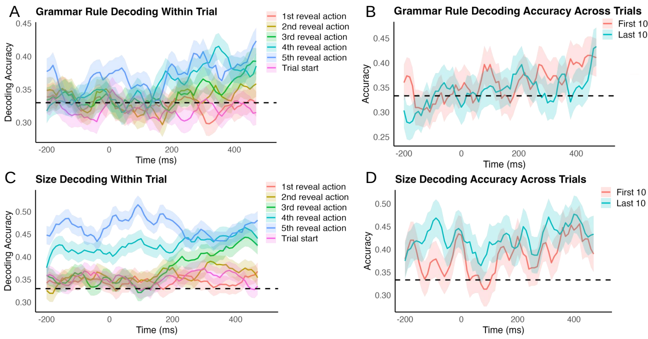

Figure 2: (A) Decoding accuracy for grammar rule classification over a 700 ms window aligned to each participant's actions (i.e., revealing a red tile). Solid lines indicate the mean decoding accuracy across participants, smoothed over time; shaded error bands represent the standard error of the mean. The dashed horizontal line indicates the chance level (33 %). Decoding accuracy for the latent grammar rule increases progressively across sequential reveal actions within a trial. Time is aligned such that 0 ms to the trial start (when the grid is displayed). (B) Grammar rule decoding accuracy after the third reveal action, separated by trial progression. The red line shows the mean decoding for the first ten trials; the blue line shows the mean decoding for the last ten trials. (C) Decoding accuracy for generative size classification within trials. (D) Decoding accuracy for generative size across trials. Format and conventions follow panels A and B.

To examine how decoding accuracy evolves at each step within a trial, we analyzed performance separately for each reveal action. We found that decoding accuracy for grammar rule identity increased within trials as participants revealed more tiles (Fig. 2A–B), with accuracy rising above chance. Early in the trial, decoding was not reliable: the first reveal was significantly below chance ($M = 0.319$, $t = -2.17$, $p = .0305$), and reveals 1–3 were not significantly different from chance ($p > .35$). Accuracy improved starting at the fourth reveal ($M = 0.351$, $t = 2.98$, $p = .0031$) and became highly significant by the fifth ($M = 0.369$, $t = 3.70$, $p = .0003$), indicating that structure becomes more decodable as information accumulates. A linear mixed-effects model confirmed this trend, showing a significant effect of reveal sequence on decoding accuracy ($B = 0.009$, 95% CI = [0.006, 0.012], $p < .0001$).

Although we also hypothesized that decoding accuracy would improve across trials, this effect was not statistically supported in our current data (Fig 2B ; $B = 0.0001$, 95% CI = [-0.0001, 0.0003], $p = .296$). Though this may be because of our modest sample size and the fact that fatigue may have limited our across-trial learning effects. Prolonged tasks can lead to performance decline, potentially masking improvement. Future work will quantify fatigue using behavioral accuracy or reaction times, and may restrict analyses to early, low-fatigue trials or shorten experimental sessions to reduce its impact.

A similar pattern emerged for decoding generative size. Using a separate decoder, we found a significant increase in accuracy within trials ($B = 0.0226$, 95% CI = [0.0189, 0.0262], $p < .0001$; Fig.2C), and a marginal improvement across trials ($B = 0.042$, 95% CI = [0.0002, 0.0841], $p = .057$; Fig.2D). While the across-trial effect did not reach conventional significance, the trend suggests participants may gradually learn size-related abstractions over time. Decoding of size was already above chance at trial onset and increased steadily thereafter. One plausible explanation is that participants could infer information about the generative size even before any tiles were revealed.

In summary, this study examined how humans acquire task-specific abstract concepts by decoding compositional structure from MEG data during learning. We can decode both the grammar rule and the generative size of the stimuli from neural signals. Decoding accuracy increased within trials, suggesting that participants incrementally constructed internal representations of the underlying structure as they gathered evidence. These results provide neural evidence for the dynamic formation of abstract concepts, supporting theories of hierarchical and compositional inference in human cognition.

References

Barack, D. L., Bakkour, A., Shohamy, D., & Salzman, C. D. (2023). Visuospatial information foraging describes search behavior in learning latent environmental features. *Scientific Reports*, *13*(1), 1126. doi: 10.1038/s41598-023-28328-3

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*(4), 954–967. doi: 10.1016/j.cell.2020.09.031

Braun, D. A., Mehring, C., & Wolpert, D. M. (2010). Structure learning in action. *Behavioural Brain Research*, *206*(2), 157–165. doi: 10.1016/j.bbr.2009.09.040

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(1), 197–209. doi: 10.1037/a0017808

Hull, C. L. (1920). Quantitative aspects of evolution of concepts: An experimental study. *Psychological Monographs*, *28*(1), i–86. doi: 10.1037/h0093130

Kumar, S. M., Dasgupta, I., Cohen, J., Daw, N., & Griffiths, T. L. (2021). Meta-learning of structured task distributions in humans and machines. In *Proceedings of the international conference on learning representations (iclr).* (URL: https://doi.org/10.48550/arXiv.2010.02317)

Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., & Behrens, T. (2023). Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. *Cell*, *186*(22), 4885–4897. doi: 10.1016/j.cell.2023.09.036

Tafazoli, S., Bouchacourt, F. M., Ardalan, A., Markov, N. T., Uchimura, M., Mattar, M. G., & Buschman, T. J. (2024). *Building compositional tasks with shared neural subspaces.* (bioRxiv preprint)

Wang, J. X. (2021). Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, *38*, 90–95. doi: 10.1016/j.cobeha.2021.01.003