Neural Computational Mechanism of Compositional Generalization



Zilu Liang

University of Oxford

St Edmund Hall

A thesis submitted in partial fulfilment of the
requirements of the University of Oxford
for the degree of

Doctor of Philosophy

Trinity Term 2025

Acknowledgements

Pursuing a degree is a long and, at times, lonely journey. What began as an exciting yet daunting step into a new field—and a new country—has now reached this point after four years of dedication. I could not have come this far without the incredible support I've received along the way.

First and foremost, I would like to thank my supervisors, Chris and Miriam, for all the guidance in science and in life. Thank you, Chris, for guiding me through confusion, for teaching me so much about how science can be communicated in a poetic and beautiful manner and for supporting me to freely explore my interest while gently keeping me on track. Thank you, Miriam, for being my biggest cheerleader throughout, for your endless patience when I needed help with even the smallest technical details, for making me feel safe about making mistakes, and for reassuring me that I would always have someone in my corner.

Thank you to the many lab members and friends who has made NRH such a wonderful space to work in Oxford. To Jess, thank you for shifting my perspective on what it means to grow as a scientist. To Mira, my wonderful bubble tea buddy, thank you for my very first hug in Oxford and for lifting me up during the harder moments. To all the collaborators I worked with on various projects—especially the data collection team that joined me in Granada—thank you. Sophie, thank you for being there from the beginning and for calming us down during chaotic times. Denis, thank you for the many bus-stop missions to find participants. Jirko, Ellie, Jacques, and Kai — collecting data in a language I barely speak was daunting; I truly couldn't have done it without you.

As someone shy and sometimes hesitant of making new connections, I never imagined building such close bonds so quickly when moving to a foreign country. To Leonie — my lovely desk mate, my loyal fMRI buddy, and my most supportive friend — thank you for all the Zilu-tailored birthday cakes, the warm hugs, and the unwavering support. Thank you, Yiran and Rainy, your friendship reminded me I wasn't alone. And to Rong, thank you for showing me that vulnerability is a strength, and for being there in quiet, meaningful ways.

Thank you to my parents, for forever believing in me and indulging my adventures. None of this would have been possible without your love and support. To Shen, my ally in science and partner in life — thank you for the fun math puzzles, for your steady presence (even from afar) through every high and low, and for helping me understand what is truly important in the grand scheme of life. To Doudou and Xuxu, my dearest friends in China, eight hours apart — thank you for always being on my side and for showing me that distance is never a barrier to sincere friendship. Growing together since undergrad has been a rare and lasting gift — no matter where our paths now lead, your presence has shaped mine. May there be many more years to come, and many more chances to meet and share warm hugs.

Lastly, I would like to thank the natural world in Oxford — all things bright and beautiful, all creatures great and small — whose quiet presence has brought me so much peace and comfort.

Content

Acknowledgements.....1

Content.....3

List of Main Figures.....6

Abstract.....8

Chapter 1 General Introduction.....9

1.1 Generalization.....10

1.2 Understanding generalization through the lens of representation 26

1.3 Abstraction — What makes a representation useful.....33

1.4 Compositionality — How representation is used for generalization.....37

1.5 The discovery of structure.....39

1.6. Thesis Outline.....43

Chapter 2 Neural representation that supports compositional generalization.....46

2.1 General Introduction.....47

2.2 The Neural Solution Space of the problem.....49

2.3 Methods.....52

2.4 Results.....68

2.5 Discussion.....84

2.6 Limitations.....88

Chapter 3 Solving compositional rule generalization task with low and high dimensional representation.....91

3.1 Introduction.....91

3.2 Experiment 1: Inducing high dimensional representation with auxiliary feature reconstruction loss.....93

3.3 Experiment 2: Inducing high dimensional representation with unsupervised pretraining.....98

3.4 Experiment 3: The flexibility of high-dimensional representation 101

3.5 Summary & Discussion.....105

Chapter 4 Flexible cross-context compositional generalization 108

4.1. Introduction.....108

4.2. General Methods.....112

4.3. Pilot M1V1: one map task with the symbol combination stimuli 118

4.4. Pilot M2V1: multi-map task, blocked evaluation.....124

4.5 Pilot M2V2 and pilot M2V3: multi-map task, randomised evaluation, new classification method.....130

4.6. Symbol representation in the cluster task.....143

4.7. Summary and Discussion.....145

Chapter 5 Transfer of structural knowledge.....147

5.1 Introduction.....147

5.2 Experiment 1. Structure learning and transfer in human and neural network.....149

5.3 Experiment 2 Hidden layer representation that supports structure transfer in neural network.....162

5.4. Summary and Discussion.....170

Chapter 6 General Discussion.....172

Appendix A: Supplementary Materials for Chapter 2.....177

A.1 Hypothesis about rule representation.....177

A.2 Participant Exclusion based on behaviour consistency.....178

A.3 Effect of multivariate noise normalisation.....179

A.4 Retrieval pattern analysis sanity checks.....181

A.5 Stimuli group decoding.....183

A.6 Testing parallel representation with cross-validated RSA.....184

A.7 Whole-brain searchlight RSA results.....185

Appendix B: Supplementary Materials for Chapter 4.....189

B.1 Procedure of pilot M2V2 and pilot M2V3.....189

B.2 Percentage of Missing trials in the 2AFC task.....190

B.3 Reclassifying stimuli as centre versus non-centre stimuli.....191

B.4 Evaluation stage performance in the multi-map 2AFC task as a function of centre versus non-centre stimuli.....195

B.5 Behavioural consistency across the treasure hunt task and the 2AFC task.....197

B.6 Transfer learning effect.....200

Appendix C: Supplementary Materials for Chapter 5.....204

References.....209

List of Main Figures

Figure 1-1. Schematic illustration function learning task involving a quadratic function.....18
Figure 1-2. Schematic illustration of a three-layer MLP.....30
Figure 2-1. fMRI Experiment Design and Procedure.....54
Figure 2-2. Treasure Hunt Task Response Accuracy and Participant Classification.....69
Figure 2-3. Low-dimensional mental map of all stimuli in V1.....71
Figure 2-4 Low-dimensional composition in V1.....72
Figure 2-5. Low-dimensional mental map of test stimuli and low-dimensional composition in superior parietal gyrus (SPG).....76
Figure 2-6 Separation of training stimuli and test stimuli in SPG.....77
Figure 2-7. High-dimensional parallel representation for the training stimuli in hippocampus.....81
Figure 2-8. High-dimensional compositional representation in vmPFC. 84
Figure 3-1.Task set up and neural network agent.....93
Figure 3-2. Behavioural performance and hidden layer representation. (A) Location prediction performance measured by distance to correct location on each axis at the final epoch for each network.....97
Figure 3-3. Neural network performance and representation geometry during treasure hunt task learning in Experiment 2.....100
Figure 3-4. Human and neural network performance in Experiment 3. 104
Figure 4-1. Example multi-map stimuli in their ground-truth locations 110
Figure 4-2. Stimuli Set and Task Pool.....115

Figure 4-3. Procedure and performance in the one-map pilot (M1V1). 122
Figure 4-4. Procedure and response map in the first multi-map pilot (pilot M2V1).....126
Figure 4-5. Behavioural performance in the first multi-map pilot (pilot M2V1) during the evaluation stage.....128
Figure 4-6 . Illustration of compositional generalization with inaccurate rule memory.....130
Figure 4-7. Model Space.....137
Figure 4-8. Winning model in cross-validation (CV) data.....140
Figure 4-9. Participant classification results.....141
Figure 4-10. Participants' performance during the evaluation stage 143
Figure 4-11. Symbol representation in the symbol arena task.....144
Figure 5-1. Task Design.....154
Figure 5-2. Structure learning and transfer performance in the 9-grid task in humans.....156
Figure 5-3. Effect of training curriculum on participants' learning and generalization behaviour.....159
Figure 5-4. Structure learning and transfer performance in the 9-grid task in humans.....161
Figure 5-5. Structure learning and transfer performance in the 16-grid task in the neural network.....164
Figure 5-6. Proposed representational geometry that supports structure transfer.....166
Figure 5-7. Observed hidden layer representations in different runs and their correlation with transfer performance in test trials in the final epoch.....169

Abstract

This thesis hopes to provide further understanding of the representation subserving compositional generalization: how reusable knowledge is acquired from experience, how it is represented in the brain, and what are the computations over the representations aid generalization in novel situations. In Chapter 1, I presented an overview of the literature and the theoretical framework adopted by this thesis. Chapter 2 introduced the task paradigm used by this thesis. This task required participants to learn how to predict locations of treasure based on coloured animals from a subset of training stimuli with feedback and generalize the learned rules to a held-out subset of stimuli without feedback. In an fMRI study, I identified the variations in the dimensionality and orthogonality of the rule representations across hippocampus, ventral medial prefrontal cortex, posterior parietal cortex and primary visual cortex that support successful compositional rule generalization. Building on these variations, Chapter 3 reported simulation studies to explore the cost and benefit of compressing the high-dimensional representations of discrete features to solve the task. Given the design limitations of the fMRI study in Chapter 2, I conducted a series of pilot experiments in Chapter 4 that gradually improved the design for extending the task to examine cross-context generalization. In this extension, I showed that participants could combine the rules learned in different temporal contexts to make novel inferences. After characterizing the properties of abstract representations and how they are put to use in a compositional fashion to solve new problems, the thesis turns to the learning process underlying the formation of abstract representation. I further extended the task design to investigate structural transfer and how some curricula facilitated faster learning and better transfer performance than others in Chapter 5. In this task, participants were taught stimuli-label associations which was governed by a hidden relational structure. Using computational simulations, I showed that abstract representation of the hidden structure is crucial for successful generalization. In a behavioural study, I showed that: 1) participants could utilize the hidden relational structure to make predictions when probed on a new set of labels; 2) training curricula that made features that were relevant for the relational structure vary

slowly in time could promote the factorised representation of structure and facilitate transfer.

Chapter 1 General Introduction

Contents

1.1 Generalization.....

1.1.1 Associative learning.....

1.1.2 Generalization in different learning tasks.....

1.1.3 Theories of generalization.....

1.2 Understanding generalization through the lens of representation.....

1.2.1 Operational definition of representation.....

1.2.2 Modelling representation learning with connectionist models.....

1.2.3 Representation learning.....

1.3 Abstraction — What makes a representation useful.....

1.3.1 A case study with categorization.....

1.3.2 Abstract representation in the brain.....

1.4 Compositionality — How representation is used for generalization.....

1.5 The discovery of structure.....

1.6. Thesis Outline.....

Our thoughts and ideas, great or small, are often built from simpler, reusable elements. This reflects the compositional nature of human cognition. This thesis explores how these reusable elements are acquired during learning, how they are represented in the brain, and how they are combined when generalizing to novel situations. In investigating the central theme of generalization and composition, this thesis hopes to not only describe the empirical observations but also shed light on the general principles. To this end, a collection of projects that span computational, neural and behavioural approaches will be presented. To establish a shared basis for understanding later

discussions, this chapter will clarify the definition of the cognitive phenomena under investigation, introduce the theoretical and computational framework that will be used to approach our problem, as well as review the current state of research in relevant domains.

The main takeaway of this chapter is that this thesis adopts, in the words of Fodor and Pylyshyn (1988), a *Representationalist* stance, aiming to understand generalization and compositionality through the lens of representation. This chapter begins with an overview of diverse generalization behaviours — ranging from simple associations to latent structures — and the historic debate between two traditions: similarity-based and rule-based accounts of generalization. Given the limitations of this dichotomy, I will introduce the representation learning framework for understanding generalization and show how it can be formalised using connectionist models. Finally, following Chomsky's three basic questions for understanding states of the mind/brain that lead to behaviour (Chomsky, 1981, 1986), I will discuss three critical questions that a Representationalist account of generalization must address:

1) What constitutes a representation that is useful for generalization?
2) How is a representation put to use for generalization?
3) How is a useful representation acquired from learning?

1.1 Generalization

Generalization is a term used widely in cognitive science to refer to behaviour that re-uses learned responses or knowledge to deal with challenges in novel situations, such as reactions to new stimuli, predictions about upcoming states, transfer of solutions to new

problems, or facilitation of future learning. It is a hallmark of adaptive behaviour in human and other intelligent beings, enabling them to navigate and thrive in an ever-changing world.

The simplest and most basic form of generalization concerns the generalization of associations: a novel stimulus evokes a response similar to that elicited by a previously associated stimulus. While the computational mechanism of how associations are learned is not the focus of this thesis, it lays the groundwork for many of the discussions that follows, such as what can be generalized and how is it acquired from experience. Therefore, this section begins with a brief description of associative learning theory. Then, I will attempt to give an overview of a multitude of generalization behaviours documented in the learning tasks predominately used in the field and discuss the overarching theoretical ideas proposed to explain generalization in these tasks.

1.1.1 Associative learning

Associative learning concerns how associations among stimuli and/or responses are formed. It encompasses a wide range of phenomenon, including the early foundational paradigms such as classical conditioning and operant conditioning as well as contemporary research on value-based learning and decision making. We will review some of the prominent ideas in associative learning and summarize them under the three-factor learning rules framework.

Hebbian learning

One of the most eminent ideas about how associations are learned was proposed by Hebb over 70 years ago, which stated that the change in synaptic connection strength between neurons should be based on repetition and causality. This gave rise to the famously

known slogan “neurons that fire together, wire together” for Hebbian learning (Hebb, 2002). A particularly influential implementation of Hebb’s theory is the spike-timing-dependent plasticity (STDP) where the change in synaptic connection strength depends on two important factors (Song et al., 2000; Magee & Grienberger, 2020). The first is the coupling of pre- and post-synaptic activity. The second is the timing of the pre- and post-synaptic activity: when pre-synaptic activity shortly precedes post-synaptic activity, the connection strength should increase; in contrast, when the post-synaptic activity precedes pre-synaptic activity, the connection should be weakened.

Error-driven learning

Another influential computational model for associative learning was developed by Rescorla and Wagner (1972) for conditioning. They aimed to: (1) describe the update of the theoretical associative strength of a stimulus, rather than its overt response probability as in the earlier models, and (2) account for how this update depends not only on the stimulus itself but also the presence of other stimuli. The latter is not the focus of the current thesis, so we will consider the reduced version for Rescorla and Wagner ‘s model (RW model) when a single stimulus is presented alone. The core of the RW model is the error-correction mechanism. Formally, when presented alone, the associative strength V_i for stimulus i is updated as follows:

$$V_i \leftarrow V_i + \alpha \beta_i \delta_i$$

where:

- α is the learning rate,
- β_i is the saliency of the stimulus,
- δ_i is the prediction error representing the discrepancy between the actual level of association (λ) and the expected level of

association. In the case of single presentation, this expectation is simply the current associative strength V_i. Therefore δ_i is defined as:

$$\delta_i = \lambda - V_i$$

As we can see, λ represents the asymptote for associative strength, serving as the target for learning. In the case of multiple stimuli being presented simultaneously, the expectation is the sum of the current associative strengths of all stimuli presented. Each stimulus may have a different saliency, accounting for the fact that different stimuli are learned at different speed. Initially proposed to explain empirical findings in classical conditioning, the RW model has been widely used to model action learning and value learning.

Later in the 1980s and 1990s, Sutton and Barto proposed the framework for modern reinforcement learning (Sutton & Barto, 2020). Under this framework, a reinforcement learning (RL) system takes actions and receives rewards from the environment. Its goal is to maximize reward in the long run. At each time step, it follows a policy that governs action selection based on the current state. The quality of an action depends on the state it leads to. Roughly speaking, due to the long-term nature of the goal, the value of a state is the expected amount of reward accumulated over a period into the future, typically measured in time steps. The temporal difference (TD) learning algorithm was proposed as an approach for a system to learn how to predict this value through interactions with the environment. The algorithm updates the value estimates using a temporal difference error (TD error). In its simplest form, which looks only one-step ahead and is known as TD(0), the algorithm updates the value of state at time t—denoted $V(S_t)$— after observing reward R_{t+1} using the following rule:

$$V(S_t) \leftarrow V(S_t) + \alpha \delta_t$$

where α is the learning rate and δ_t is the TD error defined as:

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

γ is a discount factor for future reward and is bounded between 0 and 1, reflecting a decreasing influence of distant outcomes on the present choice. We can see that in TD(0), the target for updating $V(S_t)$ is the sum of the observed reward R_{t+1} and the estimated value of the state one step ahead $V(S_{t+1})$. The formulation of TD error is very similar to the RW model outlined above, with the extension that future outcomes are taken into account. Another way to think about RW models is to frame it as a RL problem where the starting state directly leads to the end state. Although RL was not originally proposed to explain animal learning (Sutton & Barto, 2020) and is certainly capable of handling more complicated problems than learning simple associations (Silver et al., 2017), it has gained much attraction in modelling associative learning dynamics, as evidenced in both behavioural and neural data (Schultz et al., 1997; O'Doherty et al., 2003; B. Seymour et al., 2004; Starkweather & Uchida, 2021; Amo et al., 2022; Qian et al., 2025) .

As a final note, a key ingredient of RL that has not yet been introduced is *the model of the environment* that an RL agent may possess. While such a model is not strictly necessary and many RL algorithms operate without one (known as *model-free learning*) (Watkins & Dayan, 1992; Balleine et al., 2009), it offers substantial flexibility and enables more sophisticated behaviour. The learning rules described above for learning associations can also guide the refinement of the agent's world model.

The three-factor learning rule

Below, I will provide a brief summary on how associations are built based on the three-factor learning rules framework (Gerstner et al., 2018). Notably, our use of this framework in this section is highly schematized, allowing us to describe learning on a behavioural scale or on a neural implementation scale.

The basic assumption is that memory is stored in synaptic connections and that learning alters the strength of synaptic connections, i.e., *weights*. A learning rule describes how the weight change should happen. More specifically, Δw_{ij} — the amount of change in weight between a pre-synaptic (input) node j and a post-synaptic (output) node i — as a function of three factors: the pre- and post-synaptic signal (x_j and y_i), and some modulating factor (m). When applying this to the behavioural problems, input and output nodes can be construed as neurons activated by the stimuli/stimulus-response pair to be associated respectively, the synaptic strength can be construed as the associative strength.

Building on Hebb's idea, modern three-factor learning rules states that coincidences between pre-synaptic x_j and post-synaptic activity $y_i$sets one (or several) flag e_{ij} at the synapse, called *eligibility trace*. A change in this flag indicate that the weight is ready for a change. In other words, a change in eligibility trace Δe_{ij} reflects a 'candidate weight change' computed by:

$$\Delta e_{ij} = \alpha f(x_j) g(y_j) - \frac{e_{ij}}{\tau_e}$$

τ_e is a decay time constant, reflecting the timescale over which eligibility trace operates. When τ_e is sufficiently, decay in the eligibility trace is negligible.

The change in synaptic weight Δw_{ij} only happens when a third modulating factor m is non-zero, and this modulating third factor can be a global one, equally applied to all neurons, or a neuron-specific one.

$$\Delta w_{ij} = \Delta e_{ij} \times m$$

Let us reconsider the previous Hebbian learning and error-driven learning formulation under this broader framework. In the case of traditional Hebbian learning, we can set the modulating factor as a non-zero constant, therefore, a candidate weight change Δe_{ij} instantaneously leads to an actual weight change Δw_{ij}. In the case of the error-driven learning, we can set the modulating factor to be the error term δ_i. Take the R-W model as an example. We can think of the pre-synaptic neuron as one encoding a stimulus, the post-synaptic as one encoding an action. Then, the synapse strength is essentially the association strength. Taking an action when seeing the stimulus leads to an increase in the eligibility trace. Upon observing the reward, a prediction error δ_t is generated. When expectation matches the observation, the modulating factor is equal to zero , and no synaptic change occurs. However, both underprediction and overprediction produce non-zero error signals, thereby driving synaptic modifications.

1.1.2 Generalization in different learning tasks

Generalization in associative learning

The fact that associative strengths between stimuli or between stimuli and actions can generalize to novel stimuli has long been documented in the empirical studies of associative learning. As early as in Pavlov’s classical conditioning experiments, it was found that salivation in dogs conditioned to a certain stimulus (e.g. temperature

of 50 degree) could also be evoked by other stimuli (e.g. temperature of 30 degree), although in a much less degree (Pavlov, 1906; Hull, 1947). Moreover, the degree of response was modulated by the physical dissimilarity between the conditioned stimulus and the novel stimulus. This phenomenon is known as stimulus generalization, and the pattern of decreasing response with increasing dissimilarity is referred to as the generalization gradient. Both have been repeatedly documented in other early conditioning studies across multiple aspects of responses, including speed, strength, and frequency (Mednick & Freedman, 1960; R. A. Rescorla, 1976). Accumulating evidence suggest that it is quite universal, independent of the sensory modality of the stimulus, the nature of the response (learned or innate), and observed across many species (Ghirlanda & Enquist, 2003).

The discovery of generalization gradients also led to a heated debate around whether stimulus generalization reflects an adaptive process to apply previously learned associations or a mere inability to distinguish between perceptually similar stimuli. Indeed, evidence has shown that the generalization gradient is highly correlated with discrimination difficulty (Lashley & Wade, 1946; Laufer & Paz, 2012), which later contributed to the idea that the maladaptive tendency to over-generalize in some clinical disorder may stem from impaired discrimination acuity (Struyf et al., 2015). On the other hand, stimulus generalization is not *solely* driven by a failure to discriminate. Studies have found evidence of generalization using pairs of clearly discriminable stimuli (Guttman & Kalish, 1956; Norbury et al., 2018), and the pattern of generalization induced by perceptual error differs from the generalization gradient (Zaman et al., 2021), suggesting that discrimination and generalization are independent processes.

Generalization in category learning

Humans not only learn simple associations but also learn about how objects and events are grouped together, i.e., categories and concepts. Categorization is a process deeply embedded in our daily experience that guides decisions and reactions (DiCarlo & Cox, 2007; DiCarlo et al., 2012). It requires generalization from very few encounters of category members to large number of novel instances. For example, a child who has seen minivans and sedans can easily infer that trucks or limos are also cars at the first sight. Infant studies have shown that by at least about 10 months of age, and possibly earlier, infants can already form perceptual categories, including both natural categories (e.g. animals) and artificial categories (e.g. horizontal vs vertical patterns) (Sloutsky, 2018). In laboratory studies with children and adults, experimenters often devise stimuli that vary along discrete (e.g. shape) and/or continuous (e.g. size) feature dimensions and group them into artificial categories. Participants receive training on a subset of stimuli-category associations either via passive viewing or trial-and-error learning. This is then followed by a transfer stage where they are asked to classify a held-out subset of stimuli into the learned categories. The category boundary can be linear or nonlinear, and it can be defined by a single feature dimension or a combination of multiple feature dimensions (Ashby & Gott, 1988).

Generalization in function learning

While categorization concerns mapping stimuli that vary along discrete/continuous dimensions onto discrete entities, function learning concerns mapping stimuli varying along continuous dimensions onto another continuous outcome. The relationship between these two variables (e.g. stimulus x and response y) can be

linear or nonlinear, monotonic or non-monotonic. In function learning tasks, participants are usually presented with a cover story, such as predicting the speed of fire spread based on wind speed (Kalish et al., 2004). Similar to the category learning tasks, participants are trained on a subset of x–y value pair associations and are later tested on a held-out subset to assess how well they can generalize the underlying functional relationship. Low error during transfer indicates successful generalization. Transfer may involve interpolation or extrapolation, depending on the where the x-value of the held-out pair falls relative to the trained value range, that is, the upper and lower bounds of x-values encountered during training. Interpolation involves predicting y-values for x-values that fall within the trained value range whereas extrapolation involves predicting y-values for x-values that fall outside the trained value range (Figure 1-1). Typically, humans excel at learning and generalizing linear relationships (Brehmer, 1974; DeLosh et al., 1997) albeit with a tendency to underestimate y-values during extrapolation (Kwantes & Neal, 2006). In comparison, nonlinear functions are more difficult to learn (Carroll, 1963; Brehmer, 1974), especially periodic ones such as trigonometric functions (Bott & Heit, 2004; Kalish, 2013).

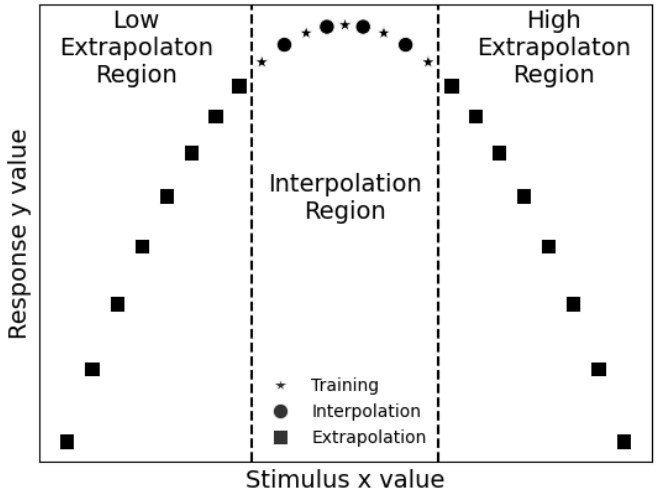


Figure 1-1. Schematic illustration function learning task involving a quadratic function. Individual points indicate x-y value pairs used for training (stars), interpolation test (circles) and extrapolation test (square).

Generalization beyond superficial similarity

In the cases we have examined thus far, training and generalization stimuli often share superficial physical similarities. In category learning tasks, they generally exhibit overlapping or similar perceptual features, whereas in function learning tasks, they vary along a single continuous dimension. A critical question then arises: can human also generalize among situations that are not superficially similar?

The answer is yes. Quite early on, researchers observed that when two superficially distinct stimuli A and B are separately paired with the same outcome or antecedent, human participants, and animals as well, indirectly learn the association between A and B. This phenomenon, known as acquired equivalence, further enables the generalization of additional knowledge from stimulus A to stimulus B (Hall et al., 1993; Honey & Hall, 1989; Meeter et al., 2009). This has latter been extended into the reinforcement learning tasks where stimuli A and B are not separately paired with a concrete stimulus, but have correlated reward probabilities (Wimmer et al., 2012).

More importantly, in addition to generalizing among discrete stimuli, whether based on superficial features or through learned associations, human and nonhuman primates can generalize among situations that share abstract structures. George Lakoff and Mark Johnson (Lakoff & Johnson, 2003) argue that human thought processes are largely metaphorical. Some concepts can be understood directly, especially those heavily grounded in our bodily experience, such as spatial concepts of UP-DOWN, FRONT-BACK, or IN-OUT. Yet, for most concepts, understanding comes at least partly from relating them to other concepts. Moreover, beyond our conventional conceptual system, imaginative and creative metaphors generate new understanding of our experience. Indeed, many scientific theories find their origins in analogies. For example, by drawing the analogy between electrons orbiting the nucleus and planets orbiting the sun, Niels Bohr shaped early thinking about atomic structure.

Empirical studies in psychology have investigated the process of analogical transfer, how humans use analogy from a semantically distant domain to guide the solution of novel problems. In a seminal paper, Gick and Holyoak (1980) presented a story to participants and tested if they could use the information in this story to generate analogous solution for semantically different problems. In one experimental condition, participants were first presented with a story about capturing a fortress by sending small forces along multiple roads to avoid mines on the road that could destroy the army. Later, they were asked to come up with a solution for a radiation problem where a tumour needs to be destroyed by high-intensity rays without damaging healthy tissue surrounding it. In this condition, the frequency of using the analogous solution to the radiation problem was higher compared to a control condition who were not exposed to

the story prior to solving the problem. Gick and Holyoak linked this observation to the notion of schema, i.e., knowledge structures within the memory, and argued that schemas afford the basis for analogical transfer (Gick & Holyoak, 1980, 1983). This ability to engage in analogical reasoning, identifying and transferring structural knowledge to novel problems, can be tested with more formalized paradigms such as the Raven's Progressive Matrices (Carpenter et al., 1990; Lovett & Forbus, 2017). More recently, such paradigms have also been used to develop benchmarks to measure human-like cognitive abilities in AI systems (M. Mitchell, 2021; Moskvichev et al., 2023; Webb et al., 2023).

Schemas not only guide problem solving but also facilitate future learning. In his pioneering work, Harlow (1949) challenged the traditional learning theories that focused on learning simple stimulus-outcome associations in a single task and explored how animals, as he put it, "learn to learn more efficiently". In the study, monkeys were presented with 344 object discrimination problems, each requiring them to learn to select the rewarded object from a new pair. Initially relying on trial-and-error, they gradually recognized the pattern that if one of the objects was not rewarded, the other would be. After approximately 200 problems, monkeys achieved 90% accuracy after following the first trial's feedback, demonstrating strong evidence of one-shot learning. Harlow replicated this finding in human children, who learned at a much faster rate than monkeys. Since then, multi-task learning, the phenomenon that training on multiple tasks that have similar structure accelerate learning of the specific task at hand, has then been documented widely in human and other nonhuman primates (A. B. Baram et al., 2021; Samborska et al., 2022; Zhou et al., 2021).

1.1.3 Theories of generalization

Initial theories of generalization sought to identify a singular mechanism that governs generalization behaviour. Two families of theories were proposed and pitted against each other, one of which argues that humans generalize based on similarity and the other argues that humans generalize based on rules. However, with accumulating empirical evidence, it became easy to find counter-examples that make either account fail. Subsequently, the field recognizes that generalization involves multiple interacting systems. In the next paragraphs, I will discuss how these two traditions have contributed to the understanding of generalization in various learning tasks, and how they come together in contemporary theories of generalization.

Similarity-based account of generalization

The generalization gradient observed in classical conditioning studies has motivated the early theories of generalization. These theories build on the assumption that there exists a psychological space and the distance in this psychological space scaled with the similarity of features between stimuli. The probability that response associated with a learned stimulus will generalize to a novel stimulus is a *monotonic* function of the distance between the novel stimulus and the learned stimulus in this psychological space (Shepard, 1987). The principle of similarity-based generalization was first quantitatively tested by Shepard and Chang (1963). Participants were trained to associate each of eight colours with one of two arbitrary labels, with the set evenly divided. As the theory predicted, the errors made during classification (i.e., misclassifying stimulus A into the category of stimulus B) was accounted for by the pair-wise similarities between stimuli.

The idea of similarity-based generalization gave rise to two prominent models of category learning or concept learning: the prototype model and the exemplar model. Both assume that categorization judgements about new object are based on its proximity to known category members. In the prototype model, all encountered members of a category are compressed to form a representative prototype. New objects are then compared to the protypes of each category (Osherson & Smith, 1981). On the contrary, the exemplar model posits that individual encounters are stored, and relevant encounters are retrieved and compared against the new object for assigning a category label (Nosofsky, 1986; Kruschke, 1992; Nosofsky et al., 1992).

The exemplar model was later extended to an associative learning model (ALM) for human functional learning and has been shown to effectively account for interpolation, particularly of linear functions (DeLosh et al., 1997). In this framework, stimuli are represented by a set of input nodes, each corresponding to one possible stimulus magnitude. Presentation of stimulus activates all input nodes, and the activation of each input node is scaled by its similarity to the stimulus's magnitude. Responses were modelled in a similar vein by a set of output nodes. During learning, the weights connecting input nodes to output nodes were adjusted by a delta learning rule similar to the Rescorla-Wagner model described in the previous section. As a result, similar stimuli activate overlapping input patterns and produce output responses modulated by similarity.

While generalization gradients have been widely observed in different tasks, several theoretical debates remain unresolved. One prominent issue concerns the shape of the generalization gradient, whether it is symmetric or asymmetric, and whether it follows a Gaussian, exponential, or alternative form (Ghirlanda & Enquist, 2003). Another

key challenge lies in how similarity is defined and reliably measured, which is also a central critique raised by proponents of rule-based accounts. Some adopted a geometric approach that define similarity as spatial distance in a low-dimensional psychological space recovered from behavioural data (e.g., confusion rates and reaction times in discrimination tasks or similarity ratings) using statistical techniques like multidimensional scaling or clustering (Shepard, 1980). This resulting psychological space is a *metric space,* meaning that distances—interpreted as similarities—must satisfy four mathematical axioms: identity (distance to self is zero), positivity, symmetry, and the triangle inequality. However, Tversky demonstrated that empirical similarity data often violate symmetry and the triangle inequality and proposed an alternative approach that defines similarity based on feature-matching (Tversky & Shafir, 2004). In this account, similarity is a weighted function of shared and distinctive features, which can be viewed as an extension of the Hamming distance from information theory. This view was later challenged by Nosofsky (1992), who argued that similarity should be understood as component of cognitive process, and that apparent asymmetries can arise from symmetric similarity measures combined with biased decision processes.

Rule-based account of generalization

As we can see, a lot of the early evidence of similarity-based generalization refers to the similarity of superficial perceptual features. This process is thought to be non-deliberative (Wills, Milton, Longmore, Hester, & Robinson, 2013). In contrast, rule-based generalization emphasizes the importance of deliberation and analytic process during problem solving. The debate between similarity-based and rule-based may be most heated in the domain of

categorization behaviour. Murphy and Medin (1985) argued that the psychological space over which similarity is defined is highly context-dependent. Some features may be relevant in one context but maybe irrelevant in another. They further argued that our judgment of similarity may depend on categorization rather than the other way around, challenging Tversky's restriction of similarity to perceptual process (Medin et al., 1993). The rule-based theorists (Murphy & Medin, 1985; Smith et al., 1992) posit that learners identify a set of relevant features and form "mini-theories" about the nature of categories (e.g. 'birds are feathered creatures that have wings and lay eggs'). Categorization of a novel object is then achieved by intentionally applying these mini-theories (e.g. 'Bats have wings, but they don't lay eggs, so they are not birds').

A parallel distinction exists in function learning. Similarity-based models such as ALM are considered non-parametric models because no explicit functional forms are learned. Alternatively, rule-based models are known as parametric models because they propose that human test hypotheses about the underlying functional form between stimuli and response (Brehmer, 1974). These models formalize the learning process as the optimization parameters β_1, β_2,... β_n in an n-th order polynomial function $y=\sum_{i=0}^{n} \beta_i x^i$ that maps stimulus x to output response y. Other variants of this model apply a log transformation to the input and output values before entering them into the polynomial (Koh & Meyer, 1991). To explain why some functions are more difficult to learn than others, the rule-based theorists have proposed a hierarchy of hypotheses: a learner sample the hypotheses about functional forms in a particular order, potentially influenced by the frequency of functions in their environment (Carroll, 1963; Brehmer, 1974). In addition, Koh and Meyer (1991) proposed that learners apply Occam's Razor, aiming to identify the simplest function that

adequately fits the data, i.e., a trade-off between parsimony and accuracy. Formally, this approach estimates the order of the polynomial n, along with other parameters such as the learning rate and the accuracy-parsimony-trade-off parameter, based on participants' training data (Koh & Meyer, 1991; DeLosh et al., 1997). For example, a fitted n=1 indicates that participants approximated the training data with a linear function, reflecting the linearity bias in function learning.

Hybrid models and converging traditions

Increasing evidence has shown failures of singular mechanisms to fully explain patterns of generalization in category and function learning, which has motivated a shift to hybrid models that integrate both similarity-based and rule-based strategies (category learning: Smith & Sloman, 1994; Erickson & Kruschke, 1998; Pettine et al., 2023; function learning: DeLosh et al., 1997; Kalish et al., 2004; Mcdaniel & Busemeyer, 2005). Many factors are proposed as the determinants underlying strategy shifts. These shifts may depend on stages of learning (Sloutsky, 2003), how the artificial categories are constructed by the experimenter (Smith & Sloman, 1994), and/or task demands (Ashby & Maddox, 2005).

Moreover, hybrid models using Bayesian principles have argued against this dichotomy between rule and similarity. Instead of modelling them with separate mechanisms, Bayesian models for concept learning and function learning have established the duality between similarity-based and rule-based generalization (Tenenbaum & Griffiths, 2001; Goodman et al., 2008; Kemp & Tenenbaum, 2008; Lucas et al., 2015). In this framework, hypotheses are generative models of the data (e.g. rules defining a category or the mapping function between input and output), and learning is conceptualized as

a problem of determining how much a learner should believe in some hypothesis *h* after some observation. This belief is described by $P(h|d)$, the posterior distribution over an a priori-defined hypothesis space *H* ($h \in H$) based on some observed data *d*. When the posterior is sharply concentrated around a single hypothesis, behaviour appears deterministic and rule-like. One example of such scenario is when there is diagnostic or sufficient data, an ideal Bayesian learner converges on the correct hypothesis. In contrast, when data is sparse, the posterior distribution is more diffuse, resulting in behaviour that appears similarity-based. Notably, a sharp posterior need not imply correctness, it may also arise from strong inductive biases (Xu & Tenenbaum, 2007).

Finally, the variety of generalization behaviour that humans and other animals exhibit can hardly be reduced to a single rule-based or similarity-based process. For instance, in analogical transfer and multi-task learning, applying learned principles to new problems requires the recognizing structural commonalities between past and new problems (Gentner, 1983; Holyoak & Koh, 1987). While this recognition can be seen as a form of similarity-based processing, it extends beyond simple perceptual similarities that traditional similarity theories typically address, echoing the long-standing criticism from the rule-based theorists that the psychological space over which similarity is defined remained elusive. Relatedly, in the viewpoint proposed by Pothos (2005), a rule-based process can be considered to be a similarity-based process where only a small number of object properties are involved to define the psychological space. Therefore, the seemingly disparate theoretical perspectives represent the two extremes in the same continuum of similarity operations. Taken together, these insights motivate our shift in focus toward one crucial question at the heart of this thesis: if objects and

events are encoded within a psychological space in which similarity is computed, what properties of this space make it effective for generalization, and how is such a space developed through learning? In other words, we turn to the fundamental question of representation.

1.2 Understanding generalization through the lens of representation

Representation is a broad term that refers to the neural signals encoding information, from internal states to external stimulations. Researchers focusing on understanding cognition from the perspective of representations are known as *Representationalist* (Fodor & Pylyshyn, 1988), and classically define cognition as computation over representations to produce behaviour (Barack & Krakauer, 2021). As a machinery for processing information, the human brain is often thought to demonstrate hierarchical and modular organization (Meunier et al., 2010; Suzuki et al., 2023): information is processed through parallel nodes and transformed as it progresses through successive stages of the processing hierarchy. Therefore, understanding the representation within a region gives insight into the computation it performs at its level of the hierarchy and the functional role it plays in the cognitive process.

In this section, I will first focus on a fundamental methodological question, what does it mean to have a representation of something, i.e., how does neuroscience research operationally define representation. Next, I will introduce the connectionist models this thesis will use to implement computational principles and generate predictions about representation. Finally, I will conclude on the

representation learning framework, and the three critical questions that it needs to address.

1.2.1 Operational definition of representation

As building blocks of the brain, neurons are thought of as the fundamental units for representing information. In the early days, neurophysiology experiments have often been interested in identifying the parameters in the experiment that drive changes in the activity of neurons. It could be the feature of sensory stimuli (e.g. the orientation of the gratings), a certain aspect of behaviour (e.g. velocity of movement), or the response rule of a task (e.g. match or no-match). When such a correspondence between a parameter and the activity of neurons is established in the empirical data, researcher would claim that they found representations of the parameter. This has been extended to the understanding of cortical functions. Cognitive neuroscience research often looks for the regions where the activity (measured using blood-oxygen-level-dependent signal in functional MRI or event-related-potential components in EEG) correlated with experiment parameters. There have been quite a few successes under this traditional approach. It can be traced back to Hubel and Wiesel's ground-breaking work on feature detection in cat's visual cortex where they observed cells that were activated by orientations of slit of the (Hubel & Wiesel, 1962). Another example is the discovery of place cells (O'Keefe & Dostrovsky, 1971), hippocampal neurons that fire in a single or multiple locations in an environment as animals freely explore the environment. These successes provide support the notion of selective tuning: neuron's activity is related to a single parameter of the experiment.

However, accumulating evidence has found that many neurons demonstrate mixed selectivity, they simultaneously encode multiple

parameters in the experiments (Rigotti et al., 2010; Warden & Miller, 2010). How is information represented if we can no longer boil single neuron's activity down to a specific experimental parameter? As it turns out, information can still be reliably decoded from the population activity even after we remove the pure selective neurons from the population (Rigotti et al., 2013). Parallel observation was found at a mesoscale with functional MRI, Haxby et al., (2001) reported that after removing voxels that responded selectively to object categories, information can still be reliably decoded from the ventral temporal cortex. Together, the evidence speaks to the idea that information is represented in a distributed fashion by the neuron population.

This perspective shift — from a neuron doctrine, which emphasizes the function of individual neurons, to a population doctrine that focuses on distributed representations and population coding (Saxena & Cunningham, 2019; Ebitz & Hayden, 2021) — has subsequently led to a methodological shift in how we analysis and interpret neural data (Chung & Abbott, 2021). For example, hypothesis-driven methods such as decoding or representational similarity analysis (Kriegeskorte, 2008; Diedrichsen & Kriegeskorte, 2017; Bernardi et al., 2020), are employed to test whether specific parameters of the experiment are being represented, and more importantly examine the relationship among those representations. More recently, researchers have adopted geometric and dynamical systems approaches that seek to systematically characterize the structure of the neural state space — the space defined by the activity of a population of neurons — and relate its geometry back to parameters of the experiment. While the neural state space is high-dimensional (depending on the number of neurons), correlations in neural activity often allow it to be approximated by a low-dimensional manifold. Understanding the

dimensionality of this manifold and how it corresponds to parameters of the experiment (Altan et al., 2021; Jazayeri & Ostojic, 2021) offers a bottom-up, data-driven approach to studying neural representations. Moreover, incorporating time as an additional dimension reveals how neural population dynamics unfold during behaviour. At any given moment in time, the state of the population occupies some point in the neural state space, and as an animal performs a task (e.g. making a decision), these points form a trajectory in this space over time. The evolution of the trajectory often provides insights into the computations carried out in the neural dynamics (Vyas et al., 2020). Finally, the geometric properties of the neural manifold have been shown — either in principle using theoretical models (Wakhloo et al., 2023; Pagan et al., 2025) , or empirically (Schneider et al., 2023; Chou et al., 2024; Fascianelli et al., 2024) — to be directly related to behaviour.

1.2.2 Modelling representation learning with connectionist models

Given our theoretical stance, we need an approach that models cognitive processes in a way that leads to testable predictions about the neural representations. One family of model particularly well-suited for this purpose is the connectionist model. Inspired by the biological brain, early connectionists build a class of computational models called neural network that consist of interconnecting, neuron-like units that operates in parallel. Therefore, these models inherently represent information in a distributed fashion. They have since evolved into deep learning , a flourishing branch in machine learning

with the potential of recreating or even exceeding the powerful capabilities of biological brains.

One canonical form of neural networks is the feedforward network, which consists of layers of units connected only in a forward direction. The number of layers[1] is known as the *depth* of the network. Each layer comprises varying number of individual units that operate in parallel and independently of one another. This architecture is also referred to as a multilayer perceptron (MLP). In an MLP, the activation of each unit is computed as a function of the weighted sum of activations from the preceding layer. This function is called the activation function. Figure 1-2 shows a sketch of a simple three-layer MLP with different activation functions. The weights effectively reflect the strength of connections between units, analogous to synaptic connections between biological neurons. Crucially, neural networks do not store explicit representations in the form of unit activations. Instead, they learn by adjusting their weights through an error-driven learning process, typically using backpropagation. Other more advanced architectures of neural network can be derived from MLPs. For instance, allowing the network to receive its state at the previous timestep as part of its input at the current timestep makes it a recurrent neural network (RNN), an architecture used widely for sequential tasks and used widely as model of the cortex in working memory task (Masse et al., 2019). Large-scale artificial intelligence systems can also be built by scaling the width or depth of the network, or by combing multiple modules with different architectures.

[1] Some researchers use the word ‘layer’ to describe layers of modifiable weights instead of layers of units. Therefore, a three-layer network in this thesis may be described as two-layer network under different definition.

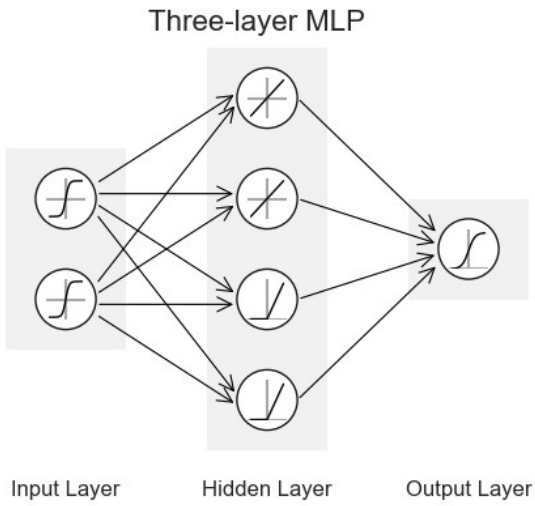


Figure 1-2. Schematic illustration of a three-layer MLP. Shown in each unit is its effective activation function: all units in the input layer use the hyperbolic tangent (tanh) activation function; the top two units in the hidden layer have linear activation, whereas the bottom two are rectified linear units (ReLU) that yield zero activation for negative values; all units in the output layer use a sigmoid activation function.

Indeed, ever since its invention, neural network models have earned much success in modelling the various types of learning behaviour, including those described in the previous section. In fact, many of the similarity-based and rule-based accounts discussed earlier have been implemented through neural network models. However, they often focused more on how representations are *used* to solve new problems rather than how these representations *emerge* through learning. One example is ALCOVE (attention learning covering map), an exemplar-based category learning model proposed by Kruschke (1992). ALCOVE is a feedforward architecture with three layers. Stimuli are assumed to come from a psychological space and each input node of ALCOVE encodes one dimension of this space (e.g. size, shape, luminance etc). Importantly, unlike general purpose MLP whose hidden layer

units has no pre-defined representational roles, ALCOVE's hidden layer consists of units explicitly setup to represent exemplars. The exemplar hidden units are then connected to the output layer with units encoding category labels. Each hidden unit's activation depends on the psychological similarity between the input stimulus and the exemplar it encodes, with attentional strengths modulating the contribution of each psychological dimension to similarity computations. During learning, the attention strengths and the association weights connecting exemplars to category labels are all adjusted using error-driven mechanisms. Therefore, ALCOVE explains why certain psychological dimensions contributes more to similarity-judgements, and how this can be biased by the examples seen during training.

Even without enforcing such strong constraints on the representational roles of the hidden units, networks trained end-to-end, analogous to how human participants are trained to perform task, can still capture key characteristics of human behaviour and neural representations when they incorporate biologically inspired architectural features. A prominent example are convolutional neural networks (CNNs), a family of feedforward networks used widely in computer vision. CNNs incorporate several biologically inspired features into its architecture, such as nonlinear activation functions and divisive normalisation (Yamins & DiCarlo, 2016). Covering evidence has demonstrated remarkable correspondence between representations of layers of CNN and that of the primate ventral visual stream responsible for object recognition. For example, the middle layers of the model are highly predictive of V4 neural responses and the output layer predictive of neural responses in the inferior temporal cortex (Yamins et al., 2014). This success demonstrates that even relatively generic learning systems, when

equipped with appropriate inductive biases, can give rise to representational hierarchies that closely resemble those found in the brain, making it a useful tool for testing hypotheses about the principles of neural computation and representation in biological systems (Kriegeskorte, 2015).

1.2.3 Representation learning

If we think of learning as an optimization process (Hennig et al., 2021), a key challenge faced by any intelligent system, artificial or biological, is how to form useful representations from often times sparse and noisy training data, whether through supervised or unsupervised learning. These representations need to support a variety of computations required by downstream tasks, such as classification, prediction, and planning. More importantly, they should not only be useful for immediate task demands but also support flexible behaviour in the long term. This challenge lies at the heart of representation learning, a growing field in machine learning (Bengio et al., 2013) that also provides a useful perspective for understanding how biological agents acquire knowledge and generalize from limited experience (Radulescu et al., 2021).

Taken together, we now have a unifying framework that enables meaningful comparisons between biological learners and computational models (especially connectionist models). It is true that this is not the solution to all the research agenda in cognitive science. One persistent issue is the lack of a formal link between representation and the biophysical properties of neuronal ensembles (e.g. neuronal-level or circuit-level constraints). While there have been some attempts to bridge the link between the tuning properties of neurons and the geometric properties of the population codes

(Kriegeskorte & Wei, 2021), this mapping remains incomplete. Therefore, in Marr's (1982) three levels of analysis—computational, algorithmic, and implementational—the approach we are using may not directly inform the implementational level. However, it offers valuable tools for the computational level (what is the goal of the system?) and the algorithmic level (what kind of representations and processes are used?). In the next three sections, I aim to review current evidence on the three questions specified in the beginning of this introduction: what makes a representation useful for generalization, how is it used, and how is it acquired.

1.3 Abstraction — What makes a representation useful

For knowledge to be transferrable across individual episodes of experience, it must be *abstracted* away from the specific details of individual episodes of experience. In other words, abstraction is the prerequisite of generalization.

1.3.1 A case study with categorization

To see why abstraction is the prerequisite of generalization, let us first consider the relatively well-studied visual pathway for solving the problem of categorization (Grill-Spector & Weiner, 2014). Exemplars within the same natural category can vary tremendously in visual appearance, yet we can reliably categorize objects within a fraction of a second (S. Thorpe et al., 1996; Liu et al., 2009; S. J. Thorpe, 2009; Sharan et al., 2014). From the perspective of representation learning, we can think of a representational space where all objects are encoded in a high-dimensional feature[2] space. Categorization (e.g.,

[2] We use the term features broadly to encompass a wide range of properties, spanning from low-level visual attributes (e.g., contrast, brightness, orientation), to

“is this a car?”) then involves learning a decision boundary (a hyperplane) that separates car exemplars from non-car ones. The most efficient representational space would therefore be one where such decision boundaries can be implemented linearly, the simplest readout mechanisms. Crucially, this decision boundary must be robust to idiosyncratic variations such as viewpoint, illumination, size, or location. Therefore, representing the object in the pixel space as received by the retina would be brittle, as pixel-level representations are sensitive to the idiosyncratic variables irrelevant for categorization.

An intuitive (but flawed) solution is to assume that each neuron could code for a specific category (i.e., "grandmother cells" or "concept cells"). This would allow for as many linear decision boundaries to be drawn as the number of neurons. While some evidence supports the existence of category-selective neurons (Quiroga et al., 2005), it is not the sole representational solution to fast and reliable categorization as this is neither scalable nor efficient: the number of possible categories is essentially unbounded, and many categories are hierarchically structured or share overlapping features. Instead, humans often reply on a modest set of core features (Biederman, 1987; Barsalou, 1999) that are often at the intermediate level of complexity, meaning that these features are neither too primitive (as in the pixel space), nor too holistic. These features act as composable building blocks from which complex categories are built. In representation space, this corresponds to a transformation from pixel space into a more abstract feature space, where these intermediate features are identified and disentangled, while the irrelevant variations are suppressed. This ‘untangling’ transformation has been empirically observed in the visual ventral stream of both humans and

mid-level perceptual features (e.g., colour, shape), and high-level conceptual properties (e.g., animacy)

macaques (DiCarlo & Cox, 2007). In early visual areas, objects are represented in a highly curved and tangled manifold, such that object identity cannot be easily readout by a linear classifier. However, as signals process through the ventral stream and reach the inferior temporal cortex (IT), the object manifold becomes flattened, enabling linearly separable representations of categories. Moreover, while famously known for showing category selectivity (such as the faces and places), IT still maintains sensitivity to certain mid-level visual features such as shapes (Brincat & Connor, 2004). Research in macaques shows that IT represents complex objects as combination of features (Tanaka, 1996; Tsunoda et al., 2001), and the encoding of these primitive features seems to operate in parallel within IT and contribute to the formation of behaviourally relevant percepts (Conway, 2018). This may underlie the observed hierarchical structure of category representations in IT (Kriegeskorte et al., 2008). Finally, the improved capacity to encode objects in an invariant fashion along the visual ventral pathway is associated with changes in the geometric properties of population activity: a decrease in dimensionality and radius of the object manifold (Froudarakis et al., 2020; Chou et al., 2024). These changes reflect a long-standing hypothesis that the brain transforms redundant sensory input into a more compact and efficient neural code (Barlow, 2012).

1.3.2 Abstract representation in the brain

From the example of categorization, we can see the key features of abstract representations:

1) **Factorisation**: relevant, contributing features are selectively identified.
2) **Compression**: irrelevant features are suppressed or discarded.

3) the resulting abstract representation is generalizable, reusable, and **compositional**.

Here, we define feature relevance in terms of its role in the generative process that produces data observed by a learner. For example, colour is not a relevant feature for distinguishing car versus non-car exemplars, but it is a relevant feature for distinguishing blueberries from other berries. These representational demands are not unique to categorization but lie at the heart of learning generalizable knowledge — i.e., latent structures that can be reused across diverse contexts. Remarkably, humans and other primates are highly adept at this, often without explicit instruction. The corresponding abstract representation of such latent structure has been reported in medial prefrontal cortex and medial temporal cortex across a variety of tasks.

One line of evidence comes from the acquisition of transition structure underlying sequential events. Using the artificial grammar learning paradigm, Reber (1967) showed that participants exposed to sequences of letters generated by a hidden grammar can later distinguish grammatical from ungrammatical sequences, despite there being no explicit knowledge of the grammatic rules. Since then, a substantial body of research has shown that primates can detect such transition structures in a stream of symbols presented in various modalities, e.g. visual, auditory, syntactic, and spatial patterns, and generalize the detected structures to novel sequences (Perruchet & Pacton, 2006; Santolin & Saffran, 2018), even across modality (Pesnot Lerousseau & Summerfield, 2024). Neural evidence suggests that abstract representations of these transition structures are encoded in the medial temporal lobe (MTL). Early neuroimaging evidence have found that MTL showed stronger activation for structured sequences (Turk-Browne et al., 2009) and that stimulus pairs with stronger

transition probabilities were represented more similarly in MTL (Schapiro et al., 2012). More recently, intracranial recordings in human patients revealed that the transition structure defining the sequence of stimuli presentation can be recovered from the population activities of hippocampal–entorhinal neurons (Tacikowski et al., 2024). Parallel support comes from model-based reinforcement learning, where the agent builds an internal model of the environment which encodes the transition probability between states in the environment (Sutton & Barto, 2020). This internal model supports the formation of a 'cognitive map' (Tolman, 1948)—a structured representation that applies not only to physical space but also to abstract, non-spatial domains (Behrens et al., 2018). Neural evidence indicates that these maps are encoded in the hippocampus and prefrontal cortex and are recruited during planning and decision-making (Wikenheiser & Schoenbaum, 2016; Niv, 2019).

Another line of evidence comes from acquisition of shared task structure underlying multiple tasks. This research builds on the foundational principles established in classical meta-learning studies by Harlow (1949) and/or analogical transfer studies by Gick and Holyoak (1980), in which human or animal participants are exposed to structurally similar problems presented with different surface features. The underlying structure—often referred to as a schema—is abstracted across these experiences. In other words, the biological learners develop representations that factorise the shared structure from the contextual details such as the surface features of individual problems. Hippocampus and prefrontal cortex (PFC) have been found to involve in such a process (Gilboa & Marlatte, 2017). For instance, Zhou et al. (2021) trained rats on a series of odour-sequence problems, in which the animals had to make a go/no-go decision based on the identity of the odour and its position within the

sequence. After learning the initial problems, rats were exposed to new problems with the same sequence and reward structure but used entirely different sets of odours. Behaviourally, rats exhibited accelerated learning across successive problems, echoing classical findings from Harlow's studies in monkeys (Harlow, 1949). Neurally, the dimensionality of representation decreased as learning progressed, leading to the emergence of a low-dimensional manifold representing the common structure of problems in orbitofrontal cortex. A similar finding was reported by Samborska et al. (2022), who trained mice on a series of probabilistic reversal learning tasks. Their results revealed distinct roles for the prefrontal cortex and hippocampus: while the prefrontal cortex encoded the abstract problem schema, the hippocampus bound this schema to the current task context, enabling flexible adaptation to specific problem instances.

In summary, this section highlights that there are two key processes that underlie abstraction: (1) the learner must identify and retain features that are relevant to the structure while discarding irrelevant or idiosyncratic features, and (2) the learner must disentangle the multiple latent factors that contribute to observable variations. This corresponds to two transformations in the representation space that support such abstraction: (1) compression, which reduces the dimensionality of input by discarding irrelevant features, and (2) factorisation, which disentangles the underlying source of variation into distinct dimensions in the representational space. Together, these transformations give rise to more abstract and structured representations useful for generalization.

1.4 Compositionality — How representation is used for generalization

Thus far, we have seen abundant evidence that abstract representations subserve generalization and allow for flexibility. This naturally leads to a central question: how are abstract representations dynamically bound and recombined to support flexible cognition? I will discuss two operations that are used to bind representations for generalization: an additive mechanism, and a multiplicative mechanism.

In vector-based natural language processing models, vector addition is widely used to as mechanisms for semantic compositions (J. Mitchell & Lapata, 2008). In such models, each word is represented as a high-dimensional vector in an embedding space. The meaning of a phrase is computed by superimposing the words in the phrase. For example, the representation of “red boat” is formed by adding the vectors for “red” and “boat”. Similar operations also underlie analogical reasoning. For instance, “queen” can be approximated as “king” − “man” + “woman” (Haga et al., 2025). There has been some neural evidence in line with this claim. Neuroimaging studies of ‘adjective + noun’ phrase comprehension (Baron et al., 2010; Baron & Osherson, 2011) found that medial prefrontal cortex (mPFC) shows an additive code for complex concept (e.g. young man as young + man). In non-linguistic domain, the additive code in mPFC has also been proposed to explain how the representation of a compound is related to its constituents, including compound rules formed by combining multiple single rules (Ito et al., 2022; Reverberi et al., 2012), values of novel food options (tea jelly) composed from its ingredients (tea and jelly)(Barron et al., 2013), complex LEGO configurations composed from simple LEGO blocks (Schwartenbeck et al., 2023), or social situations encoded as the sum of their action affordances (Thornton & Tamir, 2024). The additive operation assumes a naïve independence among the words and does not take into account sequential effects or

syntactic components. Indeed, the additive signal in mPFC has been proposed to reflect the retrieval of multiple memories maintained in an independent form (Frankland & Greene, 2020).

In contrast, the multiplicative algorithm models the composition of representations through the product of constituents. This approach allows for role-filler binding, that is, scenarios in which specific elements (fillers) need to be linked to particular functions or positions (roles) within a broader structure. Although neural evidence for multiplicative coding is limited, some findings support its role in semantic integration (K. K. Chang et al., 2009). For instance, Baron & Osherson (2011) found that while mPFC activity for adjective-noun phrases reflects additive coding, the left anterior temporal lobe exhibits patterns more consistent with multiplicative interaction (young man = young × man). More recently, theoretical work by Bakermans et al. (2025) leverages outer product operations to yield conjunctive representations of structure and content in the hippocampus.

Both operations have been widely adopted by connectionist models. Standard units in neural network models usually compute a weighted sum of inputs followed by a nonlinear activation function, implementing vector addition. In contrast, some models use multiplicative units, whose activation is a function of the products of inputs, as part of the architecture (Smolensky, 1990). Despite their theoretical appeal, relatively little empirical evidence has directly tested the presence of such operations in the neural representations of biological agents. It is likely that the former is related to situations where multiple (independent) pieces of information need to be integrated to make inference, whereas the latter is related to relational reasoning or structure binding, where interactions between inputs are critical for representing complex dependencies in a task.

1.5 The discovery of structure

Perhaps the most useful way of gaining abstract, reusable knowledge is via direct instruction, much like how children are taught the distinctions between mammals and reptiles in school. In the context of analogical transfer, Gick and Holyoak (1983) compared different ways to promote a learner's ability to identify the latent structure underlying a problem, i.e., develop a problem schema, that can be transferred to semantically different problems. When paired with comparative examples, instructions such as summarization, diagrams, or verbal principles significantly facilitated the formation of problem schemas in the learners and promoted the transfer of solution from the examples a structurally similar but semantically different problem. Understanding how to design and deliver instructions that effectively promote the transfer of skills remains an active area of research in education.

While latent knowledge can be taught directly via instruction, most of the evidence reviewed so far showed that humans (and other primates) excel at discovering the latent structure without such explicit guidance. A recent study showed that while discovery of latent structure can be achieved via different means, the resulting representation does not depend on how the structure was acquired (Courellis et al., 2023). In this study, intracranial recordings in hippocampus were collected from patients performing a reversal learning task. The latent task structure was that there were two fixed stimulus-action maps applied to the same set of stimuli. Unsignalled switches between the two mappings occurred at random points during the experiment, requiring participants to infer if a switch occurred based solely on the feedback. Therefore, successful response in the first trial after a switch indicated successful inference

and discovery of the latent structure. Crucially, patients performed multiple sessions and were provided verbal instructions on the latent task structure after the first session. The study found that patients who only understood the latent structure after instruction developed abstract hippocampal representation of latent structure similar to those who had inferred the rule without instruction in the first session. Conversely, participants who failed to grasp the latent structure, even after being told explicitly, did not develop such abstract representations in the hippocampus.

Why do abstract representation of latent structure emerge without explicit requirement to do so? And how do biological learners discover structure without explicit instruction?

A prevailing view is that sensory representations capture the statistical regularities of the natural environment (Simoncelli & Olshausen, 2001). Sensory systems adapt, through evolution, development and experience, to the statistical structure of their inputs, leading to efficient coding. This process reduces redundancy in sensory signals and maximizes information transmission. This idea is formalized under the Free Energy Principle (Friston, 2010), which proposes a unifying theory of brain function: motivated by the need to maintain its own stability in the face of an ever-changing environment, biological agents continuously minimizes "free energy," a function of sensory input and a probabilistic representation of its causes. This means that through interaction with the environment, an agent optimises its internal models about the generative process that produces input it receives. Under this framework, perception, action, and learning all serve the same goal—minimizing uncertainty about the world through active inference.

Consequently, the type of statistical structure that neural systems are optimized for varies along the sensory processing hierarchy. Early sensory areas tend to be tuned to the statistical structure of low-level, local features such as spatial, temporal, and spectral information (Baden et al., 2020). In comparison, higher-order areas are often found to encode mid- to high-level perceptual and conceptual features that are maximally useful for classification (Ullman et al., 2002). For instance, an fMRI study in humans has found that representations of objects in parahippocampus captures the statistical regularities of object co-occurrence in visual scenes and language (Bonner & Epstein, 2021). Similarly, hippocampal representations are thought to compress sensory experiences into more efficient memory codes , as supported by auto-encoder style models of hippocampal function (Gluck & Myers, 1993; Spens & Burgess, 2024; Kerrén et al., 2025). Related theoretical work has shown that when constraints such as non-negativity and energy minimization are imposed on neural activity, abstract representations capturing the data-generative structure of the input emerge as the optimal solutions (Whittington et al., 2022).

A key mechanism leveraged by neural systems to learn latent structure efficiently is temporal continuity. In sensory cortices, this is thought to underlie the development of an object representation that affords perceptual constancy (DiCarlo & Cox, 2007; DiCarlo et al., 2012). For example, when a car moves in front of you, within a short amount of time, the retinal image evolves rapidly, yet its identity remains unchanged. This temporal evolution thus indicates which image changes are identity-preserving transformations. Supporting this idea, research has shown that if object identity was consistently swapped across size or position changes during exposure, IT neurons altered their size/location tolerance (Li & DiCarlo, 2008, 2010, 2012).

These findings are consistent with Hebbian synaptic plasticity mechanisms, which are capable of extracting stable components from temporally structured input (Oja, 1982, 1992) and extracting slowly varying features from high-dimensional input streams (Wiskott & Sejnowski, 2002). These principles have been extended to model how humans rely on temporal context to form orthogonalized task representations to prevent interference and allows for efficient task switching (Bouchacourt et al., 2020). Such mechanisms based on temporal continuity may also explain the varying effectiveness of different curricula in inductive category learning in which participants have to infer the category boundaries from exemplars without explicit instruction. (Carvalho & Goldstone, 2014, 2015, 2017; Zaki & Salmi, 2019; Abel et al., 2021). Behavioural and eye-tracking studies have shown that a blocked training curriculum—where all exemplars from one category are presented consecutively—helps learners encode the shared within-category features more effectively. In contrast, an interleaved curriculum, where exemplars from different categories are mixed, facilitates better category discrimination. From the perspective of temporal learning of slow varying features, in blocked training, the shared within-category features are the slow-varying features, whereas in interleaved training, the diagnostic features that distinguish categories become the slow-varying features. Thus, both training schedules may exploit temporal continuity but do so in different ways depending on which structure—within-category similarity or between-category contrast—is emphasized.

Learning also operates across multiple timescales. The Complementary Learning Systems (CLS) theory proposes a division of labor between fast-learning systems like the hippocampus and slower-learning cortical areas (McClelland et al., 1995; Kumaran et al., 2016). The later leads to the factorisation of shared structure from

varying implementation contexts. Evidence in mice (Richards et al., 2014) shows that mice trained on multiple water-maze platforms learned the underlying pattern shared across the patterns. Over multiple days, mice showed increased sensitivity to new platforms that conflicted with the pattern. Notably, this ability was disrupted by inhibition of the medial prefrontal cortex (mPFC), highlighting its role in extracting or representing the structure over experiences spanning multiple days. Similarly, Baram et al., (2024) found that in humans, the representation of shared task structure was absent shortly after learning but emerged over the timescale of days. In artificial intelligence, a similar idea is implemented in meta-reinforcement learning (meta-RL). A meta-RL agent consists of an outer loop that optimises slowly across multiple tasks and an inner loop that adapts and optimise more rapidly for the current task demand (J. X. Wang et al., 2018; Botvinick et al., 2019). This enables the agent to learn new tasks much faster without extensive training on the new task, replicating the behavioural pattern observed by Harlow (J. X. Wang et al., 2018). Experimental evidence from mice suggest that computations in the orbitofrontal cortex support such meta-learning processes (Hattori et al., 2023).

1.6. Thesis Outline

This thesis is interested in the questions of compositional generalization: how reusable knowledge is acquired from experience, how it is represented in the brain, and how the computations over the representations aid generalization in novel situations.

In the literature review, we highlight that forming abstract representations is crucial for such behaviour. Two core aspects of the abstraction are: 1) factorisation — identifying different factors and

contributing to the data-generative process and 2) compression — reducing input dimensionality by discarding irrelevant features. Previous work in compositional generalization has largely focused on functional learning, which involves mappings between continuous input and output features (Kalish et al., 2004; Barron et al., 2013; Bongioanni et al., 2021). As the features in these tasks are continuous, the resulting representations of the features would already be along a one-dimensional continuous axis, making these tasks less suitable for addressing questions about compression. Moreover, studies directly probing the computations underlying composition remain relatively sparse. To address these limitations, this thesis leverages an established task (Dekker et al., 2022) which requires the learning of two distinct mappings, each projecting a set of discrete features onto a one-dimensional continuous axis, taught simultaneously. This setup offers a testbed for examining the three key feature of abstract representation highlighted in this introduction: 1) factorisation — how representations separate out the two rules; 2) compression — whether high-dimensional representations of discrete features are reduced to a low-dimensional 1D format after learning; and 3) compositionality — how representations of the two mappings are combined to support inferences about novel feature combinations. Chapter 2 presents an fMRI study that investigated the neural representation underlying successful compositional generalization behaviour in this task. We focused on the variations in neural geometry across key brain regions involved in the generalization process. In particular, we compared the degree of compression across regions and tested the vector addition hypothesis in each region. Based on the empirical findings, we conducted a simulation study in Chapter 3 in which we used neural network as computational models to explore the cost and benefit of compressing

the high-dimensional representations of discrete features to solve our task. Given the design limitations of the fMRI study in Chapter 2, Chapter 4 presents a series of pilot experiments that gradually improved the design for extending the task to examine cross-context generalization. In this extension, we asked if participants could combine the rules learned in different temporal contexts to make inference about novel stimuli.

After characterizing the properties of abstract representations and how they are put to use in a compositional fashion to solve new problems, the thesis turns to the learning process underlying the formation of abstract representation. As discussed in the literature review, the discovery of common structure is promoted by exposure to multiple experiences with different surface properties and may rely on temporal regularities such as slow varying features. In light of these findings, Chapter 5 presents a study that further extended the task design to investigate structural transfer and how some curricula facilitated faster learning and better transfer performance than others. The task involved learning a mapping from discrete features to discrete labels with hidden relational structure among stimuli. We asked: 1) if participants could utilize the hidden relational structure to make predictions when probed on a new set of labels; 2) if training curricula that made relational structure more salient within a short period of time could promote the factorised representation of structure and facilitate transfer.

Chapter 2 Neural representation that supports compositional generalization

Contents

2.1 General Introduction.....

2.2 The Neural Solution Space of the problem.....

2.3 Methods.....

2.3.1 Participants.....

2.3.2 Design.....

2.3.3 Procedure.....

2.3.4 fMRI data collection.....

2.3.5 Participant classification and exclusion.....

2.3.6 Preprocessing of fMRI data.....

2.3.7 Univariate Analyses of fMRI data.....

2.3.8 Preparing for Multivariate Analyses of fMRI data.....

2.3.9 Cross-validated dimensionality estimation.....

2.3.10 Quantifying the degree of training axes alignment with Parallelism score (PS).....

2.3.11 Retrieval pattern analysis: Predicting test stimulus representation from training stimuli.....

2.3.12 Decoding analysis.....

2.3.13 Representation Similarity Analysis (RSA).....

2.3.14 Group-level statistic tests for ROI analyses and whole-brain searchlight analyses.....

2.4 Results.....

2.4.1 Visual cortex represented stimuli in a low-dimensional spatial map using location code.....

2.4.2 Parietal cortex represented test stimuli in a low-dimensional spatial map using location code.....

2.4.3 Hippocampus represented rules learned from training stimuli using a high-dimensional parallel code.....

2.4.4 vmPFC showed high-dimensional compositional neural code.....

2.6 Limitations.....

2.1 General Introduction

The brain's language of thought is compositional (Fodor & Pylyshyn, 1988; Frankland & Greene, 2020). Humans communicate (infinite) complex ideas, from one hundred years of solitude to the theory of relativity, with a finite set of vocabulary. We also develop a rich repertoire of complex motor skills from motor primitives (Thoroughman & Shadmehr, 2000; Wolpert et al., 2011).

To adopt such a principal of composition, one must first acquire the simple building blocks that can be later recomposed. While some of these building blocks can be taught explicitly, e.g. learning vocabulary for a foreign language, a large part emerged incidentally in an unsupervised fashion from experience. A substantial body of research has shown that both human and nonhuman primates can detect the statistical regularities in sensory information of different modalities, e.g. visual, auditory, syntactic, and spatial patterns, and generalize the detected structure to novel sequences (Reber, 1967; Perruchet & Pacton, 2006; Santolin & Saffran, 2018), even across modality (Pesnot Lerousseau & Summerfield, 2024). The identification of commonality and shared structures across experiences is non-trivial, it ensured reusability, i.e., understanding can be generalized to new scenarios without additional learning. The idea that we form schema, a mental model that captures the regularities across (multiple) encounters facilitates has been widely supported. Human show improved encoding and retention of information that conforms to existing schema(Van Kesteren et al., 2012), and can flexibly re-use schemas to solve perceptually dissimilar problems (Gick & Holyoak, 1980, 1983; Holyoak & Koh, 1987; Luettgau et al., 2024). Similarly,

non-human primates have also been found to extract structure shared across problems and transfer this structure to solve new problems with different physical implementations efficiently (Harlow, 1949; Samborska et al., 2022; Zhou et al., 2021).

Altogether, we have a wealth of evidence at hand showing that biological agents are capable of decomposing experiences into its constituent elements. A key question then arises, namely how does the brain reformulate the representations of individual experiences to reflect that common shared structure? Factorised representation has been found to subserve flexible and fast generalization. Artificial agents trained to perform many tasks developed functionally specialized clusters involved in different cognitive processes (Yang et al., 2019). In biological agents, structures common to problems with different physical implementations is typically extracted and encoded separately from individual problem details. Such structures, often referred to as schema, have been found to be represented in hippocampus and prefrontal cortex (Gilboa & Marlatte, 2017; Zhou et al., 2021; Park et al., 2025), the evolution of which accelerates the learning of new problems (Zhou et al., 2021; Richards et al., 2014).

Once equipped with theses building blocks, we are granted the power to generate (infinite) novel compounds efficiently to understand novel sentences and to solve new problems. Such a process requires the retrieval of representation of building blocks and manipulation to construct new representation. Neurolinguistic research has uncovered the neural circuits involved in semantic compositions during language comprehension and production (Pylkkänen, 2019). When participants are processing nouns appearing in phrases (boat in "red, boat"), left anterior temporal lobe and ventromedial prefrontal cortex show heightened activity level compared to matched conditions where nouns appeared in a list (boat in "cup, boat") (Bemis & Pylkkänen,

2011). Studies using language production tasks, where participants describe cues with phrases, report similar findings (Pylkkänen et al., 2014). In non-linguistic domain, using fMRI repetition suppression (Barron et al., 2016), a study (Barron et al., 2013) reports that BOLD response to novel stimuli built from learned components (tea jelly) decreases when preceded by their components (tea or jelly). This suggests that the representation of a novel compound stimulus is built from the simpler representation of its components. Similarly, in a visual composition task where participants are asked to build a complex configuration from LEGOs (Schwartenbeck et al., 2023), medial prefrontal cortex and hippocampus maintain a conjunctive representation of building blocks in specific relational positions.

Altogether, across linguistic and non-linguistic domains, these findings suggest that representation of a compound (e.g. phrases or complex LEGO configuration) is related to its individual constituents (words and LEGO blocks). This gives rise to the second key question on how the brain binds information together. To put this in computational terms, what is the operation over the constituents that give rise to the representation of the compound? Here, we highlight the vector addition operation in which the patterns of constituents are superimposed to form the representation of the compound (Baron et al., 2010). For example, queen = king - man + female. It is used widely in the computational linguistics and machine learning literature (J. Mitchell & Lapata, 2008, 2010) that subserves binding representations of multiple pieces of information maintained in an independent form (Frankland & Greene, 2020).

With both questions in mind, our current project seeks to understand 1) the neural representation resulting from the de-composition of learned knowledge, and 2) the neural representation resulting from re-composing individual components. To do so, we leveraged an

established behavioural paradigm that incorporated these ingredients (Dekker et al., 2022). The task required prediction of treasure location based on a stimulus (coloured animal, Figure 2-1A). Unbeknownst to participants, there were two hidden rules, each mapping one set of discrete features onto a continuous spatial axis (e.g. shapes indicated positions on x-axis and colours indicated positions on y-axis). Participants were trained on a subset of stimuli with feedback on correct locations and tested on the held-out set without feedback (Figure 2-1D). Successful performance required: 1) decomposing the stimulus-location associations from training stimuli into two rules; 2) recombining the learned rules to solve novel colour-animal stimuli. Therefore, we focused on participants who successfully solved the test problems, generalizers, and asked how their neural representation solved the de-composition problem and re-composition problem.

2.2 The Neural Solution Space of the problem

To motivate our analyses, we laid out the neural solution space of our task. For the decomposition process, we asked:

1) what was the representation of each rule?

Prior to any training, the representation of the discrete features would be k-1 dimensional (k being the number of elements in each feature set, Figure S2-1A left). After training, the representation could remain high-dimensional, or it could be compressed into a low dimensional axis (Figure S2-1A right). A high-dimensional representation allows features to be well separated, enabling multiple readout rules to be applied. In contrast, a low-dimensional representation is maximally efficient for the current task but supports only a limited set of readout rules. If the mapping from features to axis locations changes

(excluding simple reversals), learning must develop a new representation that aligns with the updated rule. In summary, the dimensionality continuum reflects a trade-off between flexibility and efficiency (Fusi et al., 2016). To address this, we looked at the dimensionality of feature representation separately for each axis in the training stimuli.

2) what was the relationship between the representation of multiple rules?

In this task, the rules described how symbolic features correspond to locations on two spatial axes. Generalizers could recruit the same representation scheme for representing features and locations. If so, the representations of the two sets of features would be orthogonal to each other (Figure S2-1B left). Alternatively, generalizers could abstract away from the spatial specifics, and use a shared relational primitive, magnitude, to represent these rules as the progression along linear axes (e.g. left-ness and top-ness). If so, the two sets of features would form two parallel representation manifolds with a certain alignment scheme (top of y-axis aligned to left of x-axis in Figure S2-1B right, or top of y-axis aligned to right of x-axis). To address this, we quantified the degree of parallelism using the absolute average cosine similarity between coding directions on different axis (e.g. crab-whale vs yellow-cyan).

Taken together, there were four possible representation strategies for the rules (Figure S2-1C): high- or low-dimensional combined with orthogonal or parallel.

For the re-combination process, we asked:

3) what was the relationship between the representation of training and test stimuli that allow for compositional generalization?

When trying to work out the location of treasure given a test stimulus, the neural representation would need to retrieve information from the training stimuli. Given that the two features in our task are independent (hence the processing of colour and shape can be exchanged or happen in parallel), we made the assumption that composition in this task happened through vector addition instead of tensor multiplication in the neural space. Therefore, making the claim that the representation in a region is compositional entails two testable predictions:

i. Composition implied information sharing across training stimuli and test stimuli. A decoder trained to decode information from the representation of training stimuli should also perform well on test stimuli representation. To address this prediction, we conducted cross-condition decoding analysis for x-axis and y-axis feature separately with logistic regression decoders (Figure 2-4A).
ii. Under the vector addition assumption, test stimuli representation could be reconstructed from the linear combinations of training stimuli. Moreover, training stimuli relevant for the current test stimulus should have greater contribution. For example, for a cyan crab test stimulus at location (-2, 2), the blue crab at (-2, 0) and the cyan elephant at (0, 2) would be the relevant training stimuli. To address this prediction, we ran regressions to predict the activity pattern of a test stimulus as the linear combination of the training stimuli (Figure 2-4C). This analysis gave rise to a weight matrix, where each row described the contribution of each training stimulus to the prediction of a test stimulus. We termed this the retrieval pattern matrix. Then we compared this to two model retrieval

patterns to see whether composition happens in high or low-dimensional neural space.

2.3 Methods

2.3.1 Participants

In total, we collected two cohorts of participants with 30 participants in each cohort. We preregistered the participant classification criteria and the hypothesis before collecting the second cohort (https://osf.io/9a6r7). Participants were recruited from University of Granada and were paid a base rate of 28 euros with performance-dependent bonus of up to 10 euros. The behavioural experiments were approved by the Medical Sciences Research Ethics Committee of the University of Oxford and the neuroimaging experiment was approved by the Research Ethics Committee at the Research Centre for Mind, Brain and Behaviour at the University of Granada.

2.3.2 Design

We leveraged a previously established paradigm, the Treasure hunt task (Dekker et al., 2022), to investigate the neural mechanism supporting compositional generalization. In this task, participants were instructed to hunt treasure on an island based on cues. Cues were coloured animals. The hidden rule was that each feature of the cue (colour or shape) indicated the treasure location on the x- or y-axis in cartesian coordinates (Figure 2-1A). The mapping from a perceptual feature to the corresponding continuous spatial axis was arbitrary and randomized across participants.

In each trial (Figure 2-1D), participants first saw the cue appear in the centre of the screen for 1 second. After the cue came off, a circular

arena appeared, and participants had to move the pirate to the treasure location with four keys that corresponded to movement in the cardinal directions. The response stage was self-paced; participants can only proceed to the next screen after they had submitted a response for the current trial. To familiarise participants with the layout of controller in the scanner, we fixed the mapping from fingers to directions (index-middle-ring-pinky corresponded to left-up-down-right). In training blocks, feedback on the correct treasure location was presented for 2s. In test blocks, no feedback was provided. Importantly, participants were only exposed to 9 stimuli during training, and the remaining 16 unseen colour-shape combinations were used as test stimuli (Figure 2-1A). Therefore, participants had to make zero-shot inferences on these unseen combinations to solve the task.

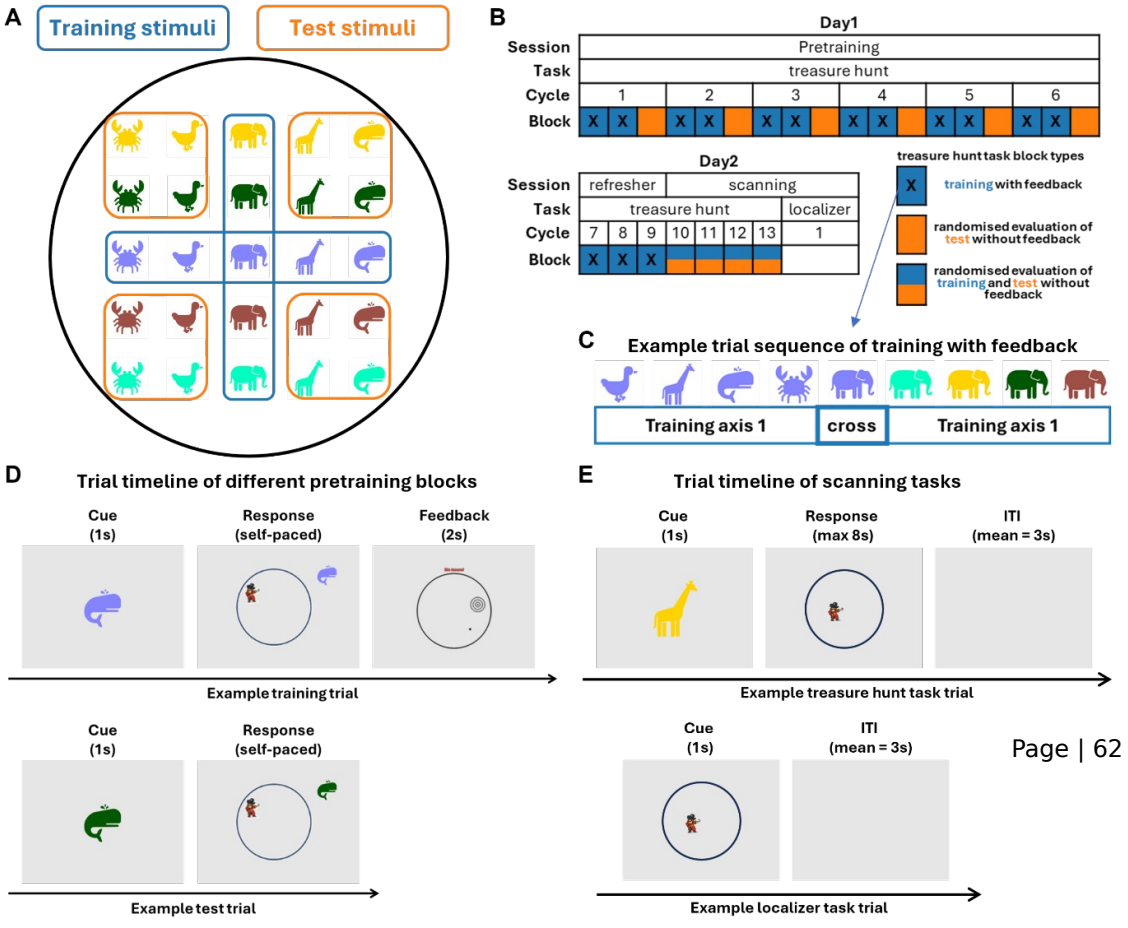
To make sure we could have as many generalizers as possible, we picked the axis-aligned blocked training curriculum from the previous study (Figure 2-1C). In this curriculum, the training stimuli locations were aligned to the cardinal axes and constituted a "+" in the centre of the map, i.e., stimuli in the middle row and column. These stimuli were presented in a blocked fashion in a training block. Participants first saw stimuli on one axis, then the stimulus at the centre location where the axes crossed, and then stimuli on the other axis. The order in which the axes were presented alternated across training blocks. In the test blocks, test stimuli were presented randomly.

During scanning, participants were evaluated on all stimuli in the treasure hunt task without feedback. The trial timeline was similar to that of the test block, except that participant were allowed a maximum of 8 seconds to make their response (Figure 2-1E top). A jittered inter-trial interval (ITI) was used (mean = 3 s, range = 1–5 s). All stimuli were presented randomly.

To obtain the neural representation of spatial locations, we designed another localizer task for scanning (Figure 2-1E bottom). In each trial, participants saw the pirate appear at a location within the arena and had to indicate whether the pirate position was the same as the previous trial using a keypress. Pirate position could be any of the nine trained locations (the centre "+" in Figure 2-1A).

2.3.3 Procedure

The study consisted of two sessions (usually two days). Participants completed a pretraining session online before coming to the refresher and scanning session in person (Figure 2-1B). In the pretraining session, participants completed 6 train-test cycles. Each cycle had two training blocks followed by a test block. In the scanning session (Figure 2-1B), participants completed a refresher of three training blocks before going into the scanner. In the scanner, participants completed four runs of the treasure hunt task followed by one run of a localizer task. Each run of the treasure hunt task had 3 repetitions per stimulus (25×3=75 trials per run). Due to time constraints,



participants were only required to move the pirate in one of the repetitions per stimulus. No feedback was provided, and all trials were randomly presented. The localizer task had 20 repetitions per training location (9×20 = 180 trials). Our analyses focused on the cue presentation stage in both tasks.

Figure 2-3. fMRI Experiment Design and Procedure. (A) Groundtruth map of the stimuli. Feedback on the correct treasure location was only provided for the nine training stimuli in the middle row and column (highlighted in blue). Participants had to infer the treasure locations of the remaining test stimuli (highlighted in orange). (B) Experiment procedure. (C) Example trial sequence in a training block. (D) Trial timeline for training block trials (top) and test block trials (bottom). (E) Trial timeline for treasure hunt task (top) and localizer task (bottom) in the scanner.

2.3.4 fMRI data collection

We acquired T2-weighted functional images on a 3T SIEMENS MAGNETOM Prisma scanner with a 64-channel head coil and a multi-band acceleration factor of 2. We acquired 50 slices, 2.5mm thick with: repetition time (TR) = 1730ms, echo time (TE) = 30ms, flip angle = 66°, field of view (FoV) = 210 mm, voxel size = 2.5×2.5×2.5 mm^3. To correct for spatial distortion, a field map was acquired with dual echo-time images covering the whole brain with the following parameters: TR = 520ms, TE1= 4.92ms, TE2 = 7.38ms, flip angle = 60°, FoV = 210mm, voxel size = 3.0×3.0×2.5 mm^3. A T1-weigted structural image was acquired with the following parameters: TR = 2250ms, TE= 4.18ms, flip angle = 7°, FoV = 256mm, voxel size = 1×1×1 mm^3.

2.3.5 Participant classification and exclusion

We classified participants based on same metric used in Dekker et al., (2022). It assumed that the response of a generalizer would follow a

binormal distribution around the true treasure location (ground truth model, Figure 2-2B). This was compared against a null model where the participant responded in any location on the island equally likely (Figure 2-2B). Log-likelihood ratio (LLR) was computed to describe whether the probability that a participant's response followed the ground-truth model $(P_{binormal})$ exceeded the probability that they were responding randomly $(P_{uniform})$.

To calculate these response probabilities, the arena was discretised into grids, and the normalized probability of response in each grid was obtained for each model. First, we defined the discretised probability mass function as the probability of response falling into each grid. For each grid in the circular arena ($x^2+y^2<2$), we calculated the unnormalized probability for each model as follows.

The uniform model was:

$$p_{raw}(x,y|uniform)=1.$$

The binormal model was:

$$p_{raw}(x,y|binormal)=\int_{ylb}^{yub}\int_{max\,(xlb,-\sqrt{2-y^2})¿}^{¿¿min\,(xub,\sqrt{2-y^2})¿¿}\Phi(x)*\Phi(y)¿$$

where xlb, xub, ylb, yub were the lower and upper bounds of the grid x,y falls in. Φ was the probability density function of the normal distribution $\Phi(a)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{(a-a_{true})^2}{2\sigma^2}}$, where a_{true} was the correct x/y coordinate, and we set $\sigma=\frac{radius_{reward}+stepsize}{3}$ (for x, y).

Next, we normalized the probabilities for each grid so that they sum up to 1:

$$p_{normed}(x,y|model)=\frac{p_{raw}(x,y|model)}{\sum p_{raw}(x,y|model)}$$

Finally, a lapse rate (λ) of 0.01 was added to allow for a small degree of randomness in the response:

$$p(x, y|model) = (1-\lambda) * p_{norm}(x, y|model) + \lambda * p_{norm}(x, y|uniform)$$

Now that we obtained the normalized probability of response in each grid for each model, the LLR of any given trial t could be calculated:

$$LLR_t = \log \frac{p(x_t, y_t|binormal)}{p(x_t, y_t|uniform)}$$

Over N trials, the average LLR was computed as:

$$LLR = \log \left\{ \prod_t \frac{p(x_t, y_t|binormal)}{p(x_t, y_t|uniform)} \right\}^{\frac{1}{N}}$$

$$¿ \frac{1}{N} \sum_t \log \frac{p(x_t, y_t|binormal)}{p(x_t, y_t|uniform)}$$

$$¿ \frac{1}{N} \sum_t LLR_t$$

$P_{binormal} > P_{uniform}$ would lead to $LLR > 0$. If the LLR of training stimuli was greater than zero ($LLR_{train} > 0$¿, the participant would be classified as a learner. If the LLR of test stimuli was greater than zero ($LLR_{test} > 0$¿, the participant would be classified as a generalizer. Importantly, we computed the LLR_{train} and LLR_{test} for each participant during their last block of pretraining and during scanning. Because there were four runs of scanning, we first computed the average response to each stimulus across all four runs for each participant, then computed LLR_{train} and LLR_{test} on the averaged response map for each participant. We classified participants based on their performance in the scanner. We further excluded 4 participants who were learners or generalizers in the last block of pertaining but behaved like non-learners or non-generalizers in the scanner ($LLR_{train} > 0$ in pretraining, $LLR_{train} < 0$ in scanner; or $LLR_{test} > 0$ in pretraining, $LLR_{test} < 0$ in scanner, see Figure S2-2). This procedure yielded 41 participants classified as learner-

generalizer (referred to generalizers hereafter) across both cohort (Figure 2-2C) and they were all included in the reported analyses. We focused on the generalizers to test our hypotheses and plotted the other non-generalizers for comparison.

2.3.6 Preprocessing of fMRI data

All fMRI data was pre-processed in SPM12. For each participant, we acquired data from 5 functional runs (EPI images), a fieldmap and a high resolution T1 image. First, we performed auto-reorientation so that all the acquired images were roughly in the same register as the MNI152NLIN6Asym template (Fonov et al., 2009, 2011). Second, we calculated the voxel displacement map (VDM) from the fieldmap image. And then, we corrected signal distortion based on the calculated VDM image and realigned the images to the first volume of the first run. Third, we coregistered the T1 image to the mean EPI image of the first functional run. Fourth, we performed segmentation and normalization on the coregistered T1 image using the default tissue probability maps in SPM. Finally, the normalization parameter estimated using the coregistered T1 image were used to normalize all the EPI images. Normalised functional images were then smoothed using a full width at half maximum that was twice the voxel size (5mm in our case).

We used the estimations of the six head motion parameters (three translation parameters on x,y,z plane and three rotation parameters: pitch, roll and yaw) generated by SPM during realignment to calculate framewise displacement (Power et al., 2012) First, we converted rotational parameters from radians to millimetres by calculating displacement on the surface of a sphere of radius 50 mm. Then, we calculated the derivatives for the six parameters (e.g. for framewise head movement on x: $d_{x_t}=|x_t-x_{t-1}|, t$ being the current timepoint).

Finally, for each TR, we defined framewise displacement as the sum of the six derivatives. If participants demonstrated excessive head motion in the scanner (more than 10% framewise displacement greater than 1 voxel size), they would be excluded from further analysis. No participants were excluded due to excessive head movement.

2.3.7 Univariate Analyses of fMRI data

For all univariate analyses, we used the general linear model (GLM) approach and conducted our analyses in using SPM12. All GLMs contained regressors for cue onset events and response onset events. In the treasure hunt task, cue onset events were modelled with a boxcar function with duration of 1s. And the response events were modelled with a boxcar function whose duration corresponded to the actual movement duration for each trial. In the localizer task, both the cue onset and the response onset events were modelled with a stick function. The design matrix was then convolved with the canonical hemodynamic response function. Each run was modelled by a separate set of regressors in the GLM with additional dummy regressors representing the run number. Motion parameter estimates and their derivatives from the pre-processing stage were included as nuisance regressors in all GLMs. We additionally scrubbed timepoints and their neighbours (from -1 and +2) with excessive head movement (where any of six derivatives was greater than one voxel size) following {Citation}Power et al., (2012). A temporal high-pass filter of 128s was applied. First-order autoregressive model was used to estimate temporal autocorrelation. An implicit threshold of 0.2 was used in combination with an explicit mask applied to limit the analysis to in-brain voxels.

2.3.8 Preparing for Multivariate Analyses of fMRI data

To understand the neural geometry that supports compositional generalization, we adopted multivariate approaches to analyse activity patterns induced by stimuli in the tasks. All multivariate analyses were conducted in Python, except for the group-level analyses for searchlight results.

We anatomically defined our regions of interest (ROIs), ventromedial prefrontal cortex (vmPFC) and hippocampus (HPC) based on previous literature suggesting their role in inference. We further included early visual cortex (V1) as a control region. vmPFC and HPC masks were generated from the AAL parcellation (Rolls et al., 2020). V1 mask was generated from the HCP-MMP1.0 parcellation (Glasser et al., 2016) that was projected onto the MNI152NLIN6Asym template (Petre et al., 2023). Before running the multivariate analyses (in ROI or searchlight), we obtained whitened activity pattern for each stimulus.

To obtain the activity pattern of each stimulus, we ran a first-level GLM to estimate the contribution of each stimulus to each voxel's activation level following the LSA approach (Mumford et al., 2012). The onset events of each stimulus in each run were modelled as a separate regressor in this GLM. We ran this GLM for each participant on their unsmoothed fMRI data in SPM12. We extracted the estimated coefficients for each stimulus as the raw (unwhitened) activity pattern matrix B where each row b_k is the unwhitened activity pattern of a stimulus in one run. Then we performed multivariate noise normalization (‘*whitening’*) on the raw activity pattern (Walther et al., 2016; Ritchie et al., 2021):

$$b_k^{\wr} = b_k \hat{\Sigma}^{\frac{-1}{2}}$$

where the voxel-by-voxel covariance matrix $\hat{\Sigma}$ is estimated from the residuals of the GLM using Oracle Approximating Shrinkage Estimator

(OAS) implemented in the scikit-learn package (Pedregosa et al., 2011). Because the covariance matrix $\hat{\Sigma}$ is Hermitian, so we can obtain the whitening matrix $\hat{\Sigma}^{\frac{-1}{2}}$ via the eigen decomposition of $\hat{\Sigma}$.

$$\hat{\Sigma}=VDV^T$$

$$\hat{\Sigma}^{\frac{-1}{2}}=V\,D^{\frac{-1}{2}}\,V^T$$

The whitened $B^{\prime}$ were used for the following multivariate analyses. The choice for adopting the noise normalisation procedure was based on the analysis of improved signal to noise ratio after noise normalization in perceptual regions (V1&V2, Figure S2-3) that were not the key inference regions of our interest.

2.3.9 Cross-validated dimensionality estimation

We estimated the functional dimensionality of fMRI data for the treasure hunt task with four-fold cross-validated singular value decomposition (SVD) (Ahlheim & Love, 2018).

To quantify the dimensionality of the representation of each rule, we examined the dimensionality of non-centre training stimuli representation on each axis separately. Let $B=\left[b_1,b_2,\ldots,b_j,\ldots\right]^T\left(b_j\in R^{n_{voxel}}\right)$ be the activity pattern matrix of m stimuli from all four runs. We split the activity pattern matrices from four runs into three subsets: fitting subset (B_{fit}, two runs), validation subset (B_{val}, one run) and test subset (B_{test}, one run). Then we took the average activity pattern of the fitting set $\overline{B}_{fit}$, and performed SVD:

$$\overline{B}_{fit}=USV^T$$

Here $U=\left[u_1,u_2,u_3,u_4\right]$ and $V=\left[v_1,v_2,\ldots v_{n_{voxel}}\right]$ were the left and right singular vectors, and the diagonal entries of S: $\left[s_{11},s_{22},s_{33},s_{44}\right]$ were the singular values. We reconstructed the data with all possible dimensionality values from 1 to m and found a dimensionality $\hat{k}$ that maximized the

correlation between the reconstructed data and the validation subset. $\widehat{\overline{B}_{fitk}}$, the reconstruction of $\overline{B}_{fit}$ with dimensionality k, was computed as:

$$\widehat{\overline{B}_{fitk}} = \sum_{i=1}^{k} s_{ii} u_i v_i^T \; where \; k = \in [1,2,\ldots,m]$$

and we computed $\hat{k}$ with

$$\hat{k} = \underset{k \in [1.2.....m]}{arg\,max} \; pearsonr\left(\widehat{\overline{B}_{fitk}}, \overline{B}_{val}\right)$$

$\hat{k}$ would be used as the estimated functional dimensionality for the held-out run. To obtain a metric for reconstruction quality on this held-out run, we computed the $\hat{k}$ dimensional reconstruction across the fitting subset and validation subset, then compared it against the held-out run. That is, we computed the average activity pattern matrix across the three runs of fitting and validation subset $\overline{B}_{fv}$, performed SVD on $\overline{B}_{fv}$, and obtained a $\hat{k}$ dimensional reconstruction of the data $\widehat{\overline{B}_{fvk}}$. We computed the correlation between $\widehat{\overline{B}_{fvk}}$and the held-out B_{test} to obtain a metric of reconstruction quality. This metric was also used as a proxy for signal quality (noise level). If there were signal present in the data, the reconstruction correlation should be greater than zero. We performed permutation test of the z-transformed reconstruction correlation at group level to test if the correlation was above zero.

2.3.10 Quantifying the degree of training axes alignment with Parallelism score (PS)

Let $X = \left[x_{-2}, x_{-1}, x_1, x_2\right] \wedge Y = \left[y_{-2}, y_{-1}, y_1, y_2\right] \left(x_j, y_k \in R^{n_{voxel}}\right)$ be the activity pattern matrix for all the non-centre training stimuli on x and y axis respectively for a single participant. Within each axis, we computed the coding direction between a pair of locations. For instance, coding direction between location i, j on x axis was $dx_{ij} = x_i - x_i$, and the same for y: $dy_{ij} = y_i - y_i$. Then we computed the cosine similarity between dx_{ij} and dy_{ij} to measure the alignment of the coding directions between i, j

on x and y axis. Finally, we obtained the parallelism score as the average cosine similarity across all possible location pairs.

$$PS = \frac{1}{C(4,2)} \sum_{i<j, i,j \in [-2,-2,1,2]} cosine\,similarity\left(dx_{ij}, dy_{ij}\right)$$

Perfectly orthogonal axes representations would lead to PS of 0 (Figure 2-7D left). Negative PS indicated that the top of y-axis was aligned to the right of x-axis, whereas positive PS indicated top-left alignment between y-axis and x-axis (Figure 2-7D right).

Due to the noisy and high-dimensional nature of fMRI data, we adopted a permutation procedure to obtain a null distribution of the PS. We generated this null distribution of PS at group level with the following procedure. We took the activity pattern matrix of all the non-centre training stimuli (each column was the activity pattern of one training stimulus, from x-2 to x2, then from y-2 to y2). For each participant, we randomly shuffled the labels of the columns (i.e., randomly assigned the activity patterns to stimuli), then computed the PS for this shuffled matrix and this participant. And then, we averaged across all participants to obtain an average PS for one shuffle. We repeated this process 10,000 times, obtained the corresponding PS values for each ROI, and used it as the group null, denoted as $PS_{HPCnull}$ etc. Finally, for visualization purpose, we fitted a gaussian distribution to the group null of each ROI.

Participants' axis alignment schemes were then classified by comparing their PS against the group null in hippocampus ROI. Let $PS_{HPC,j}$ be the PS for participant:

1) If $PS_{HPC,j}$ was significantly greater than the null ($\frac{1}{10000}\sum\left(PS_{HPCnull} > PS_{HPC,j}\right) < 0.005$), then the participant was classified as ***top-left aligned***.

2) If $PS_{HPC,j}$ was significantly smaller than the null ($\frac{1}{10000}\sum\left(PS_{HPCnull}<PS_{HPC,j}\right)<0.005$), then the participant was classified as ***top-right aligned***.
3) Otherwise, the participant was classified as ***orthogonal***.

2.3.11 Retrieval pattern analysis: Predicting test stimulus representation from training stimuli

To test the vector addition hypothesis, we ran regression to predict the activity pattern of test stimulus as a linear combination of activity patterns of training stimuli (Figure 2-4C). To avoid overfitting, we averaged the activity patterns of a given stimulus across odd runs and across even runs to obtain an odd-even split for cross-validation.

Regression weights fitted with the averaged patterns from odd split were applied to even split to validate the regression model fit and vice versa. Let $X_{tr,odd}$, $X_{tr,even}$ be the activity pattern matrix of training stimuli for odd and even splits respectively, and $X_{te,odd}$, $X_{te,even}$ denotes the activity pattern matrix of test stimuli for odd and even splits respectively. Each row of an activity pattern matrix (denoted by $x_{te,odd}$, $x_{te,even}$) is the activity pattern of a stimulus. We obtained two weight matrices W_{odd} and W_{even} in which each row (denoted by w_{odd}, w_{even}) is a set of ordinary least square estimates obtained for predicting a test stimulus ($w_{odd}, w_{even} \in R^{1\times 8}$):

$$w_{odd}=\underset{w}{arg\,min}\left\|x_{te,odd}-w\,X_{tr,odd}\right\|$$

$$w_{even}=\underset{w}{arg\,min}\left\|x_{te,even}-w\,X_{tr,even}\right\|$$

where $\|\cdot\|$ indicated L2-norm of a vector. The final retrieval pattern matrix was computed by averaging W_{odd} and W_{even}.

Model fit was evaluated on held-out set by applying the weights to make prediction about test stimuli and calculating the coefficient of determination (denoted R^2 hereafter). Note that R^2can be negative

because the held-out data were not seen during fitting. If the R^2 was significantly above zero, it indicated that the prediction in the held-out was successful. Specifically, for odd and even split, we computed the mean R^2 across prediction of all test stimuli :

$$R^2_{odd} = \frac{1}{16} \times \sum 1 - \frac{\left\| x_{te,odd} - w_{even} X_{tr,odd} \right\|}{\| x_{\quad\quad} - mean(x_{\quad\quad}) \|}$$

$$R^2_{even} = \frac{1}{16} \times \sum 1 - \frac{\left\| x_{te,even} - w_{odd} X_{tr,even} \right\|}{\| x_{\quad\quad} - mean(x_{\quad\quad}) \|}$$

The final R^2 was computed by averaging R^2_{odd} and R^2_{even}.

Moreover, the vector addition hypothesis implied that the relevant stimuli contributed to the prediction and more so than the irrelevant stimuli. Therefore, we computed the mean weights for the relevant stimuli and irrelevant stimuli separately and for each participant. Then we tested if the mean weight for relevant stimuli was greater than zero as well as greater than the mean weight for irrelevant stimuli at the group level. If the prediction was successful, and both tests were satisfied, it meant that for a given region, composition happened through vector addition.

After these sanity checks, we explored whether vector addition happened in high or low dimensional space. To this end, we first define the retrieval weight matrix for test stimuli by averaging the weights obtained from odd and even splits. Then, we designed two model retrieval patterns that corresponded to composition in high and low dimensional space and compared the obtained retrieval pattern matrix against the models. For high-dimensional composition, this process happened through perceptual feature matching, i.e., to make inference about a jade crab, one would only retrieve information from relevant training stimuli that was jade and contained crab. This resulted in a model retrieval pattern ('perceptual matching') 'here only the relevant training stimuli would have high weights (Figure 2-

4D top). If composition happened in low dimensional space, the retrieval weights irrelevant training stimuli would be further modulated by the distance to the relevant training stimuli whereby the closer distance yielded larger weights ('distance modulated', Figure 2-4D bottom). Since these two model matrices were correlated, we entered both models in the regression to predict the obtained retrieval pattern matrix to disentangle their effect.

2.3.12 Decoding analysis

All decoding analysis reported were performed using the linear logistic regression decoder with l2 penalty. To make sure the obtained results were not sensitive to random seeds, we repeated all decoding analysis with 20 random seeds and reported the average across random seeds. All decoding analysis were performed with the scikit-learn package (Pedregosa et al., 2011).

2.3.12.1 Decoding feature of training stimuli

This analysis accompanied the dimensionality analysis to address the role of dimensionality in maintaining feature separability, so we also performed the decoding analysis for non-centre training stimuli on x- and y-axis separately. To avoid overfitting to run-specific noise, we adopted the repeated stratified k-fold scheme for cross-validation. Take x-axis feature decoding as an example, for each participant, activity patterns of non-centre training stimuli on x-axis from all four runs were randomly divided into four splits disregarding the run identity. The logistic regression decoder was fitted on three out of the four splits and evaluated on the held-out split. Within the fitting data, we further performed hyper-parameter tuning with repeated stratified three-fold cross-validated grid search to look for the best performing penalty parameter. The decoder was refit to all fitting data with the

identified penalty parameter and evaluated on the held-out fourth run. This process was repeated four times where each run served as the held-out run once. Decoding accuracy was computed as the average accuracy in held-out run across the four repeats.

2.3.12.2 Cross-condition decoder generalization

To examine whether there was information sharing between training and test stimuli, we evaluated the performance of feature decoders on test stimuli after fitting the decoder using training stimuli representation and vice versa (see Figure 2-4A for schematic illustration of the procedure). We ran this two-fold cross-validated decoding analysis for x- and y-axis feature separately. Take x-axis feature decoding as an example, for each participant, we took the activity patterns of stimuli that were not in the middle column and split the data into a fitting set and an evaluation set according to stimuli condition (training or test stimuli). The logistic regression decoder was fitted with a hyper-parameter tuning procedure similar to the procedure in 2.3.12.1. After that, we applied the decoder to the evaluation set. The mean evaluation accuracy across two folds were taken as the cross-condition feature decoding accuracy for x-axis feature. Same process was repeated for y-axis feature.

2.3.12.3 Cross-task decoder generalization (from treasure hunt task to localizer task)

To examine whether the representation of stimuli in the treasure hunt task was shared with location representation, we evaluated the performance of discrete position decoders on localizer task after fitting the decoder using treasure hunt task stimuli representation. We performed the analysis for x- and y-axis feature separately and for training stimuli and test stimuli separately. The procedure was similar to that in cross-condition decoding except that the evaluation was

conducted on the location localizer task instead of the other stimuli group.

2.3.13 Representation Similarity Analysis (RSA)

In representation similarity analyses, we first calculated representation dissimilarity matrices (RDM) for the neural data and different model predictions. Then, we quantified the similarity between neural RDM and model RDMs with regression or correlation.

In terms of the neural RDM, for each searchlight sphere or ROI, we took the activity pattern matrix and calculated the neural dissimilarity between a pair of stimuli using correlational distance.

We formalized hypotheses for how stimuli are represented in the current task into different model RDMs. First, we described how each stimulus could be represented with a feature vector. Then we calculated model RDMs using Euclidean distance (for continuous features) or hamming distance (for one-hot or multi-hot encoded features) between the feature vectors of a stimuli pair. Similarity between neural RDM and model RDMs was quantified using regression or correlation analysis.

2.3.13.1 High-vs-low dimensional (feature vs location) representation of stimuli

For this analysis, we regressed neural RDM of stimuli representation on two model RDMs: feature and location. The feature RDM came from a high-dimensional representation where stimuli were transformed into two-hot encoding (colour and shape). The location RDM came from a low-dimensional representation where stimuli were transformed into their actual locations on the treasure map. We performed this analysis separately for training stimuli and test stimuli

in the treasure hunt task, as well as for the pirate positions in the localizer task.

2.3.13.2 Train-test stimuli separation

This analysis was performed on all the stimuli in the treasure hunt task. We regressed neural RDM of stimuli representation on three model RDMs: feature, location and train-test separation. The first two RDMs were derived in the same way as in 2.14.1. The train-test separation assigned a distance of zero to training-training stimuli pair and test-test stimuli pair, and a distance of 1 to training-test stimuli pair.

2.3.13.3 Parallel-vs-orthogonal representation of training stimuli

To further validate the analysis with PS, we performed a two-fold cross-validated RSA. In this analysis, the training stimuli activity patterns from two runs (fitting set) were averaged to calculate PS. Then, we ran RSA with the averaged activity pattern across the other two held-out runs (evaluation set) where we regressed the neural RDM of training stimuli on two model RDMs: parallel and orthogonal. Crucially, we only included distance between stimuli from different axis. The orthogonal RDM was generated from the distance between the ground truth location on the map. The parallel RDM was generated based on the PS from the fitting set: PS>0 indicated that the top of y-axis (y=-2) was more similar to the left of x-axis (x=-2), y=-1 was more similar to x=-1, and so on. PS<0 indicated that the top of y-axis (y=-2) was more similar to the right of x-axis (x=2), y=-1 was more similar to x=1, and so on. We repeated this process across all possible fit-evaluation splits and computed average regression coefficients across these splits as the metric of similarity with parallel or orthogonal RDM.

2.3.14 Group-level statistic tests for ROI analyses and whole-brain searchlight analyses

Due to the uneven number of generalizers and non-generalizers and low number of non-generalizers, group-level statistics in ROI were conducted with permutation test (10000 shuffles). For independent sample tests on the means, participant labels (generalizer or non-generalizer) were randomly shuffled to generate the null. For one-sample tests on the means, the sign of data was randomly flipped to generate the null. For paired sample tests on the means, data within a pair was randomly exchanged to generate the null. For correlation tests, null was generated by randomly permuting the data of one variable while the data of the other variable remained unchanged.

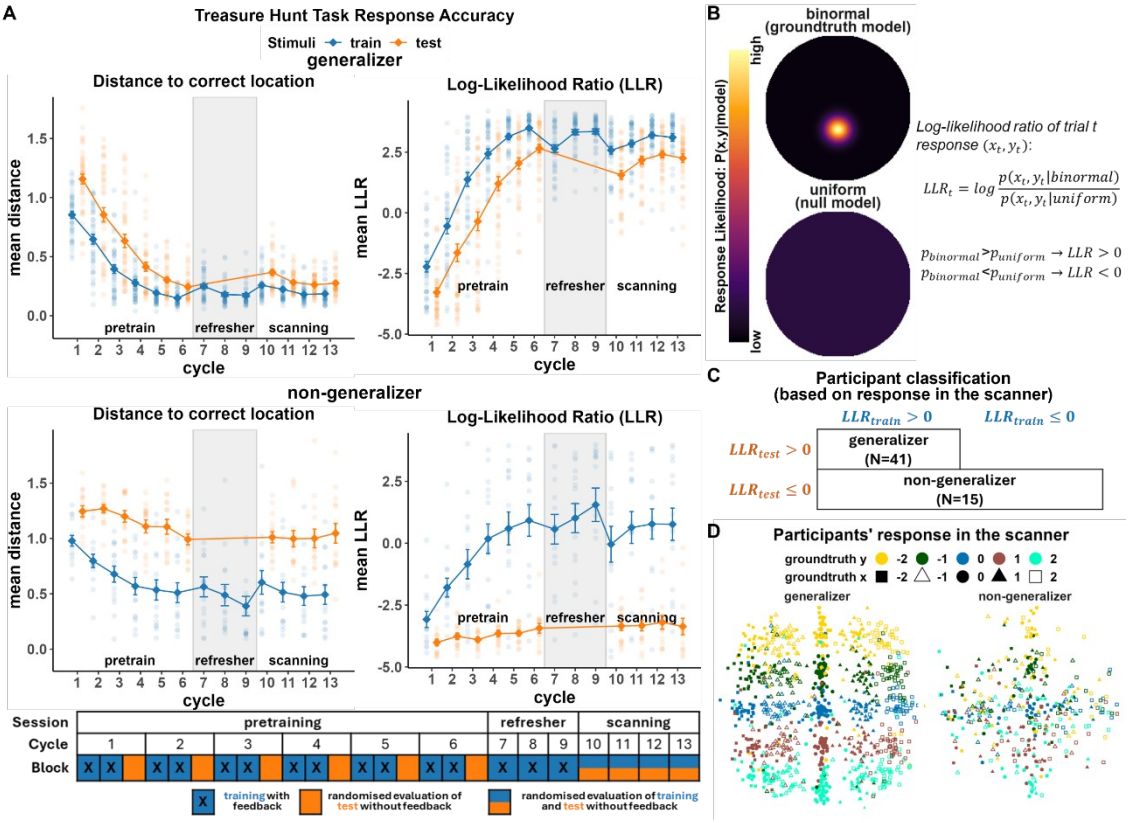
In whole-brain searchlight RSA analyses, group level analyses were only performed on the generalizers. Group-level T maps were generated using the second level analysis of one sample t-test in SPM12. We reported the clusters that satisfied either of the following criteria: 1) cluster that passed an FWE correction of p=0.05; 2) cluster with a peak that passed an FWE correction of p=0.05. Additionally, we showed the corresponding contralateral clusters even if it failed cluster correction for reference.

2.4 Results

Response accuracy of the treasure hunt task was measured with two metrics: distance to the correct location and log-likelihood ratio (LLR). Distance to the correct location was computed from the Euclidean distance between participant's response and the groundtruth location. LLR quantified whether a participant's response was better explained by a binormal model or a uniform model (Figure 2-2B). The binormal model assigned higher response likelihood near the

groundtruth location whereas the uniform model assigned equal likelihood to all locations within the arena (see Methods 2.5 for detail). Participants who successfully inferred the treasure locations of the test stimuli were referred to as *generalizers,* defined as having a mean LLR greater than 0 for test stimuli. The classification was based performance in the scanner (Figure 2-2C).

Figure 2-2A showed the performance in the treasure hunt task of generalizers and non-generalizers over the course of the experiment. Although generalizers inferred the treasure locations of test stimuli, their response accuracy was lower than for the training stimuli. The transition to the scanner initially led to a decrease in performance,



but this was recovered as participants became more familiar with the response controllers in the scanner. On average, generalizers' response in the scanner echoed the ground truth map in the treasure hunt task (Figure 2-2D). Next, we focused on the neural representation of the generalizers to address the decomposition and (re)composition question.

Figure 2-4. Treasure Hunt Task Response Accuracy and Participant Classification. (A) Response accuracy of generalizers and non-generalizers in the treasure hunt task throughout the experiment. Experiment procedure shown in the bottom for reference. Response accuracy was quantified by mean distance to correct location (left) and mean LLR (right). Smaller distance and larger LLR indicate higher accuracy of response. Solid dots and error bars indicated mean ± standard error. Opaque dots represented

performance of individual participants. (B) Heatmaps showing response likelihood when the correct location is at (0,-2) under the null (top) and ground-truth model (bottom). (C) Participant classification and final count. (D) Participants' response in the scanner. Each data point represented one participant's average response to a stimulus across 4 fMRI runs.

2.4.1 Visual cortex represented stimuli in a low-dimensional spatial map using location code

Given that generalizers accurately predicted the treasure locations for all stimuli, we first searched for the corresponding location representation in neural space. Surprisingly, the anatomically defined V1 ROI seems to strongly represent the stimuli in this low-dimensional spatial map. We performed multi-dimensional scaling (MDS) on the neural RDM of all the stimuli averaged across generalizers in V1 (Figure 2-3A). MDS of V1 (Figure 2-3B) showed a faithful reconstruction of the ground-truth map (Figure 2-1A). We conducted a whole-brain RSA searchlight to look for regions that represented all stimuli in the treasure hunt task with their corresponding locations after controlling for visual features and separation between training and test stimuli (see 2.3.13.2 Train-test stimuli separation). In line with the ROI analysis, we observed strong correlations with the location RDM in bilateral visual cortex (Figure 2-3D, Table S2-2-1A, peak coordinate [-5, -86, 10.5], $t(40) = 10.193$). A similar effect was observed in left precentral gyrus (Table S2-1A, peak coordinate [-40, -18.5, 58], $t(40) = 5.483$).

The MDS and searchlight results indicated that V1 had a low-dimensional representation. To further validate this finding, we estimated the dimensionality of feature representation in the anatomically defined V1 ROI using cross-validated singular value decomposition (CV-SVD) separately for x-axis and y-axis feature. Most

generalizers showed surprisingly low-dimensional representation of the features (Figure 2-3C).

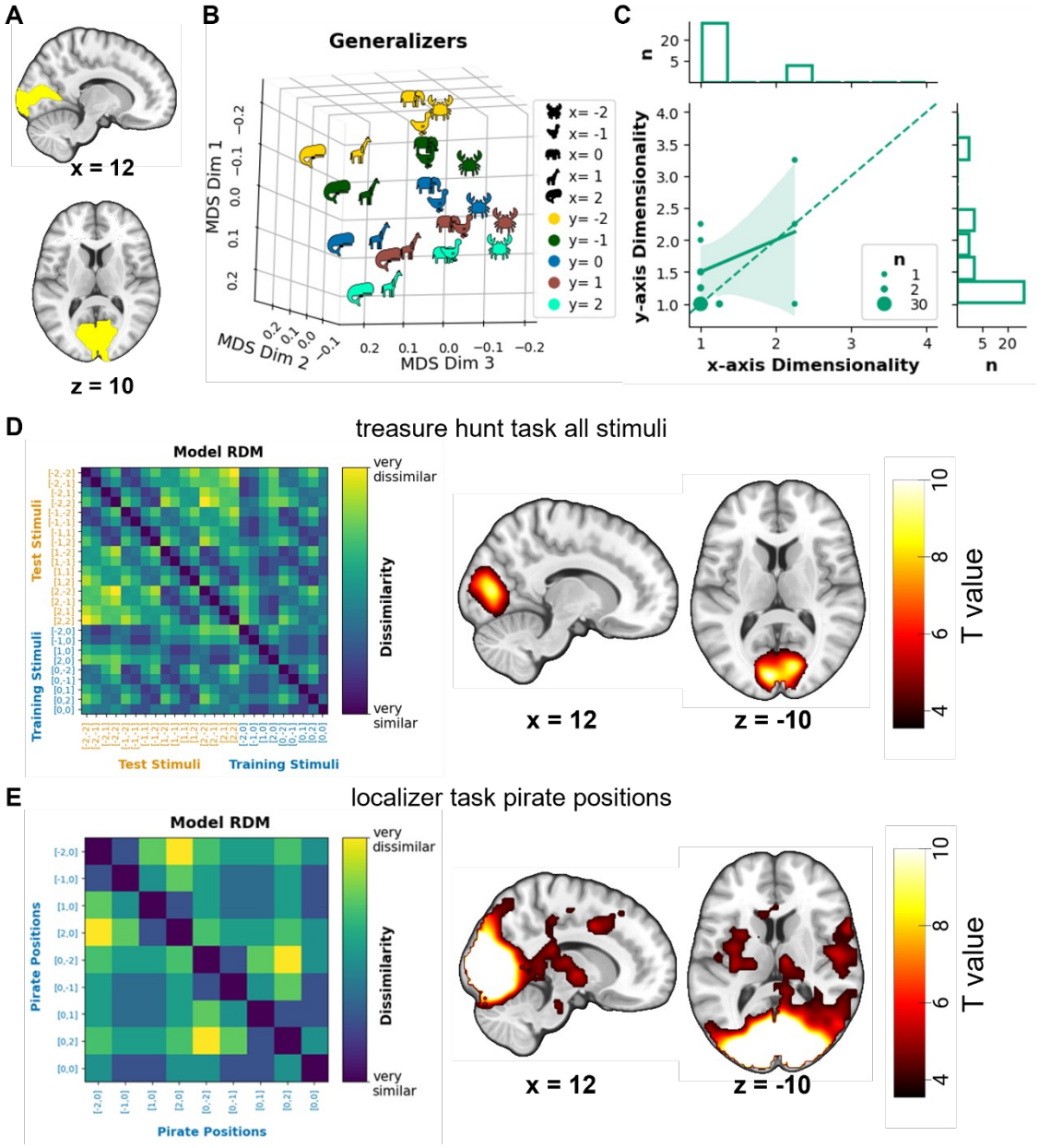


Figure 2-5. Low-dimensional mental map of all stimuli in V1. (A) Anatomical mask for V1. (B) Multidimensional scaling result from the average neural RDM of generalizers in V1. (C) Functional dimensionality of V1 ROI calculated from non-centre training stimuli. (D-E) Second-level searchlight result of location representation in treasure hunt task (D) and

localizer task for generalizers (E). Display threshold: cluster level p<0.05 FWE-corrected; cluster-defining threshold treasure hunt task p < 0.001, localizer task p<0.0001.

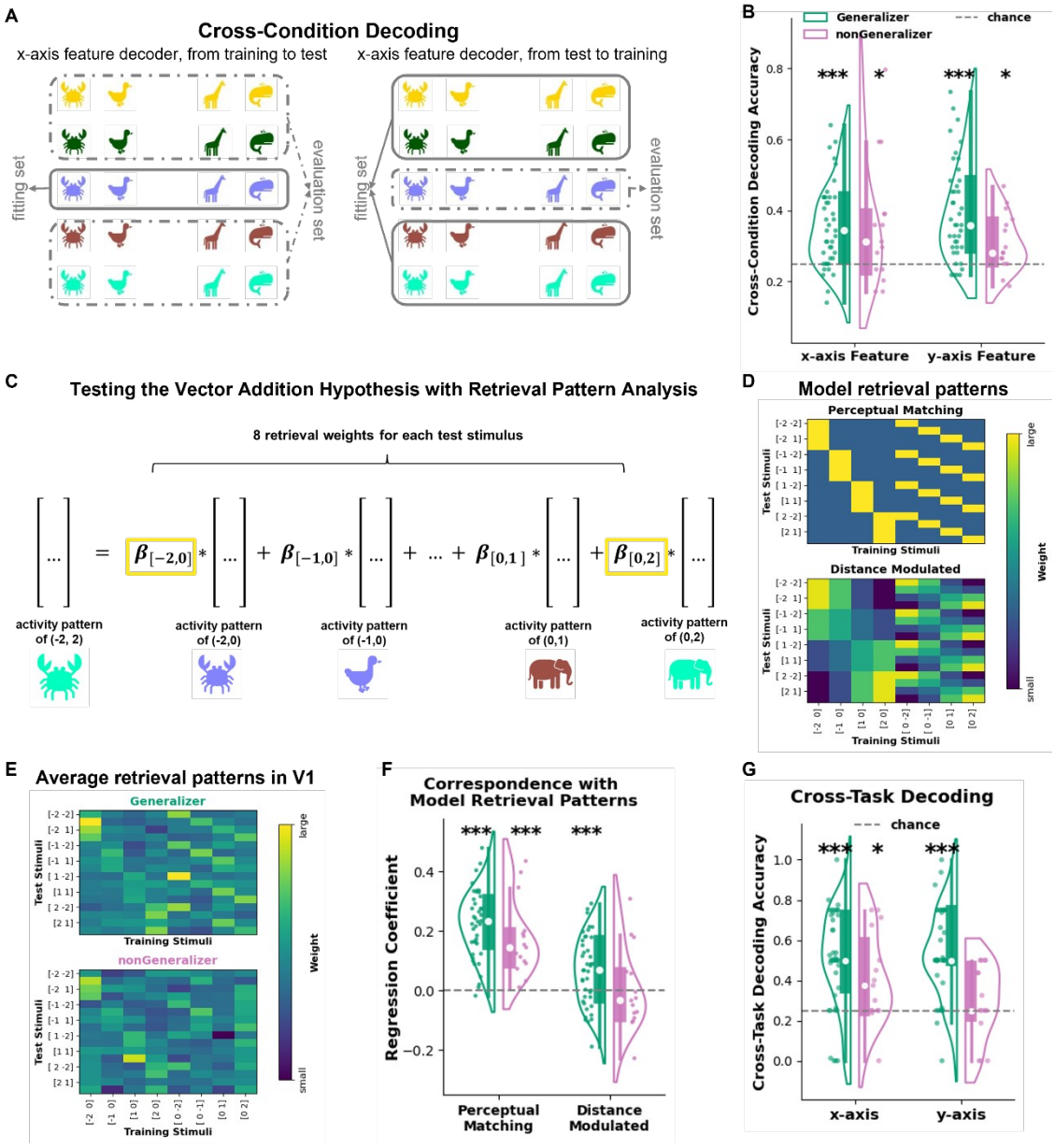


Figure 2-6 Low-dimensional composition in V1. (A) Schematic illustration of cross-condition feature decoding analysis. (B) Cross-condition feature decoding accuracy in V1 in generalizers. Dots represented the decoder's mean evaluation accuracy across 20 random seeds for one

participant. (C) Schematic illustration of the retrieval pattern analysis where the test stimulus activity pattern was predicted from a (weighted) linear combination of training stimuli activity patterns or pirate positions activity patterns. In this example, the relevant training stimuli/pirate positions at [-2,0] and [0,2] should contribute more to the prediction of test stimulus as [-2,2]. (D) Model retrieval patterns (Perceptual matching and distance modulated model). (E-F) Predicting test stimulus from training stimuli. (E) Average empirical retrieval pattern of generalizers and non-generalizers observed in V1. (F) Similarity of the empirical retrieval patterns with perceptual matching and distance modulated models. (G) Cross-condition feature decoding accuracy. Dots represented the decoder's mean evaluation accuracy across 20 random seeds for one participant. *Notes*: Asterisks indicated results of permutation tests, *p<0.05, **p<0.01, ***p<0.001. In the violin plots (B, F-G), dots represented individual participants. Box indicated the quantiles in each participant group, while whiskers extended to points within 1.5 IQRs of the lower and upper quartile.

Next, we examined whether representation in V1 was compositional and whether the composition happened in a low-dimensional space. We first looked for evidence of information sharing across training and test stimuli. Cross-condition decoding analysis (Figure 2-4A, see 2.3.12.2 Cross-condition decoder generalization) revealed that feature decoder trained on training stimuli showed significantly above-chance decoding accuracy when applied to test stimuli for both features in generalizers (Figure 2-4B). This was also true in non-generalizers, which likely reflected V1's role in feature representation. Then, we performed retrieval pattern analysis (Figure 2-4C, see 2.3.11 Retrieval pattern analysis: Predicting test stimulus representation from training stimuli) to test the vector addition hypothesis for compositional representation. Indeed, test stimuli activity patterns could be predicted from the linear combination of training stimuli patterns in V1 (generalizers: R^2 = 0.150±0.008, p<0.001; non-generalizers: R^2 = 0.128±0.010, p<0.001). The relevant training

stimuli also contributed more to the prediction than the irrelevant training stimuli (generalizers: relevant-irrelevant = 0.035±0.003, $p<0.001$; non-generalizers: relevant-irrelevant = 0.023±0.005, $p<0.001$; see Figure S2-4 for full sanity check results). We further tested if the composition happened in high or low dimensional space. If it happened in high-dimensional space, only the training stimuli with the relevant perceptual features would be retrieved to compose a test stimulus (perceptual matching model). If it happened in low-dimensional space, the retrieval pattern should also reflect the distance between training stimuli (distance modulated model). We regressed participants' retrieval pattern with these two model patterns (Figure 2-4D). The retrieval pattern in generalizers' V1 (Figure 2-4E) not only corresponded to the perceptual matching model ($\beta=0.222\pm0.018$, $p<0.001$) but also demonstrated distance modulation ($\beta=0.067\pm0.019$, $p=0.001$; Figure 2-3F). For the non-generalizers, the retrieval pattern only resembled the perceptual matching model ($\beta=0.173\pm0.033$, $p<0.001$) but not the distance modulated model ($\beta=0.001\pm0.036$, $p=0.992$; Figure 2-3F). Together, these results demonstrated that V1 representation in generalizers was compositional and low-dimensional.

The low-dimensional representation for treasure hunt task stimuli in V1 seemed to be shared with that of pirate position representation. A whole brain searchlight RSA of pirate position representation in the localizer task revealed a large cluster extending from visual cortex to parietal cortex (Table S2-3, Figure 2-3E, peak coordinate [5,-81,-2], $t(40) = 33.555$) that overlapped with the previous cluster of location representation for all treasure hunt task stimuli (Figure 2-3D). This was further supported by the cross-task decoding analysis. We found that x-/y-axis position decoders trained on activity patterns of treasure hunt task test stimuli showed significantly above-chance

decoding accuracy when applied to localizer task in generalizers (Figure 2-4G; x-axis accuracy=0.506±0.038, p<0.001; y-axis accuracy=0.410±0.058, p=0.018).

2.4.2 Parietal cortex represented test stimuli in a low-dimensional spatial map using location code

In the previous section, we conducted a searchlight analysis looking for regions that represented all stimuli with the corresponding treasure locations. This analysis inherently implied that training and test stimuli were represented in the same manifold. However, the mental processes involved in solving training stimuli and test stimuli could be different. Training stimuli trials could be solved by retrieving the correct locations from memory, while test stimuli trials would require inference from the training stimuli. This could lead to a separation in the representation space. Therefore, we performed a searchlight analysis looking for the location representation only in the test stimuli. We found a cluster extending from right precuneus to right superior parietal gyrus that represented test stimuli using their ground truth locations (Table S2-2, Figure 2-5A, peak coordinate [12.5, -66, 55.5], t(40) = 4.071). This cluster also overlapped with the previously found visual-parietal cluster that represented the pirate positions in the localizer task (Table S2-3, Figure 2-5B, Figure 2-3C red cluster), suggesting that superior parietal gyrus may use a generalized location code to represent the test stimuli and the pirate positions.

To further validate this observation, we combined the bilateral superior parietal gyrus clusters into a functional ROI and performed the dimensionality analysis and retrieval pattern analyses[3]. Indeed,

[3] The searchlight of test stimuli location representation also revealed a cluster in the left superior parietal gyrus. Although it failed to pass cluster correction, it was symmetrical to that in the right hemisphere, had stronger peak T-value compared to the right hemisphere, and had a decent cluster size of 102. Therefore, we

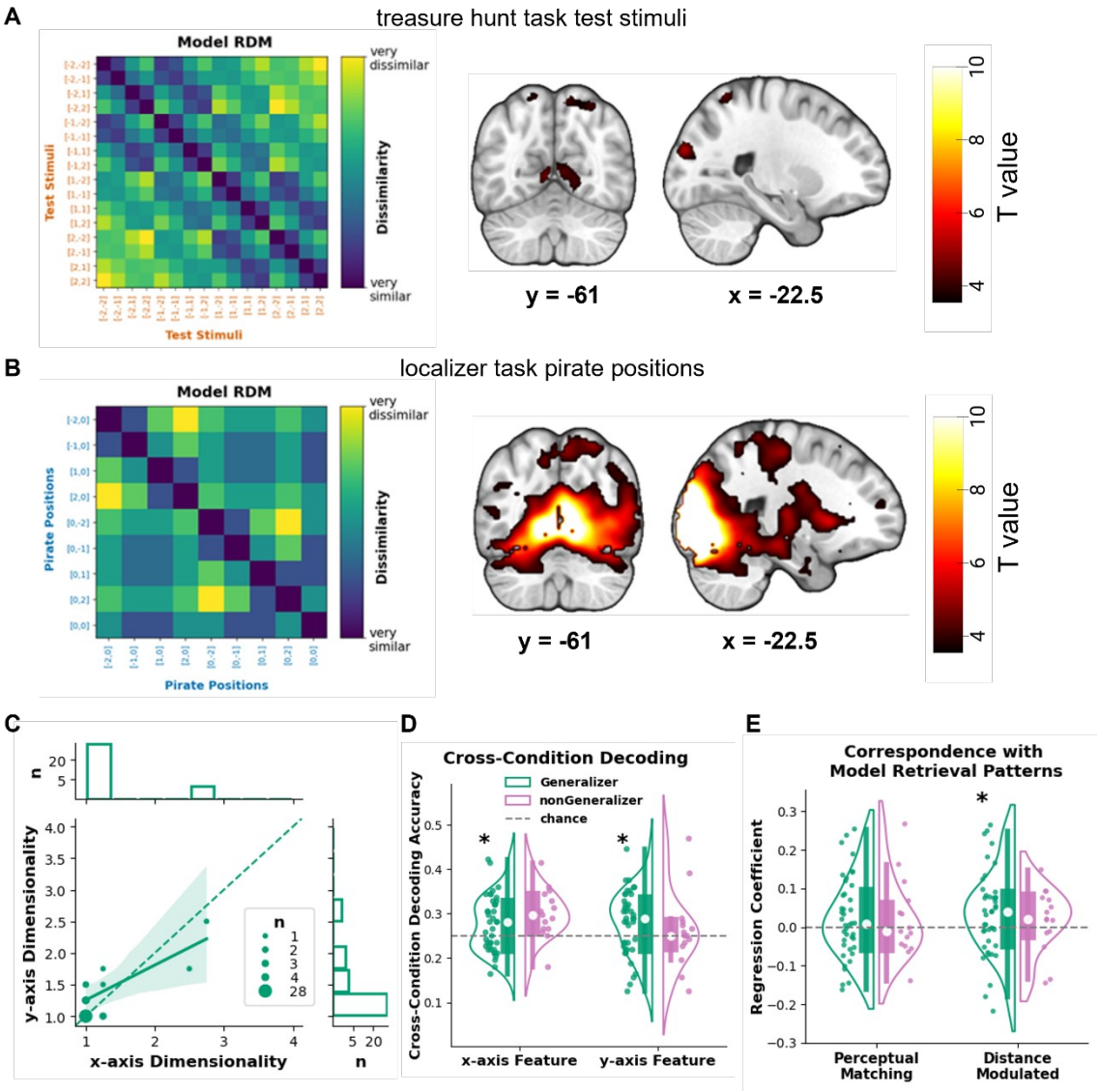
most generalizers showed low-dimensional representation of the features in the parietal ROI (Figure 2-5C). The representation in the parietal ROI was also compositional. Feature decoder trained on training stimuli showed significantly above-chance decoding accuracy when applied to test stimuli for both features in generalizers (Figure 2-5D). Representation of a test stimulus could be predicted from the linear combination of training stimuli (generalizers: R^2 = 0.097±0.010, p<0.001; non-generalizers: R^2 = 0.052±0.015, p<0.001; see Figure S2-4 for full sanity check results), and more importantly, the retrieval pattern resembled the distance modulated model (β=0.036±0.017, p=0.042) but not the perceptual matching model in generalizers (β=0.024±0.016, p=0.164;Figure 2-5E).

While also demonstrating a low-dimensional location code, this function ROI in bilateral superior parietal gyrus differed from V1 in that the training stimuli and test stimuli were represented in separate manifolds (Figure 2-6A). This train-test separation was reliably decoded from the activity patterns in the parietal ROI (generalizers: accuracy = 0.610±0.008, p<0.001; non-generalizers: accuracy = 0.628±0.019, p<0.001). Moreover, this separation was not purely due to feature difference between the training and test stimuli. This ROI was within a large parietal cluster revealed by an RSA searchlight looking for train-test separation while controlling feature representation (Table S2-1B, Figure 2-6C). To understand the cause of this separation, we visualized the neural representational similarity with location RDM in different types of stimuli: training stimuli from the treasure hunt task, test stimuli from the treasure hunt task, and pirate positions from the localizer task. We extracted the beta weights of location RDMs in different whole-brain searchlight analyses (see 2.3.13.1 in 2.3.13 Representation Similarity Analysis (RSA)) and

considered the activation to be bilateral rather than unilateral and used the bilateral ROI hereafter.

plotted the average beta weight within this function ROI for each participant. It seemed that the training stimuli were not represented as their ground truth location (Figure 2-6B).

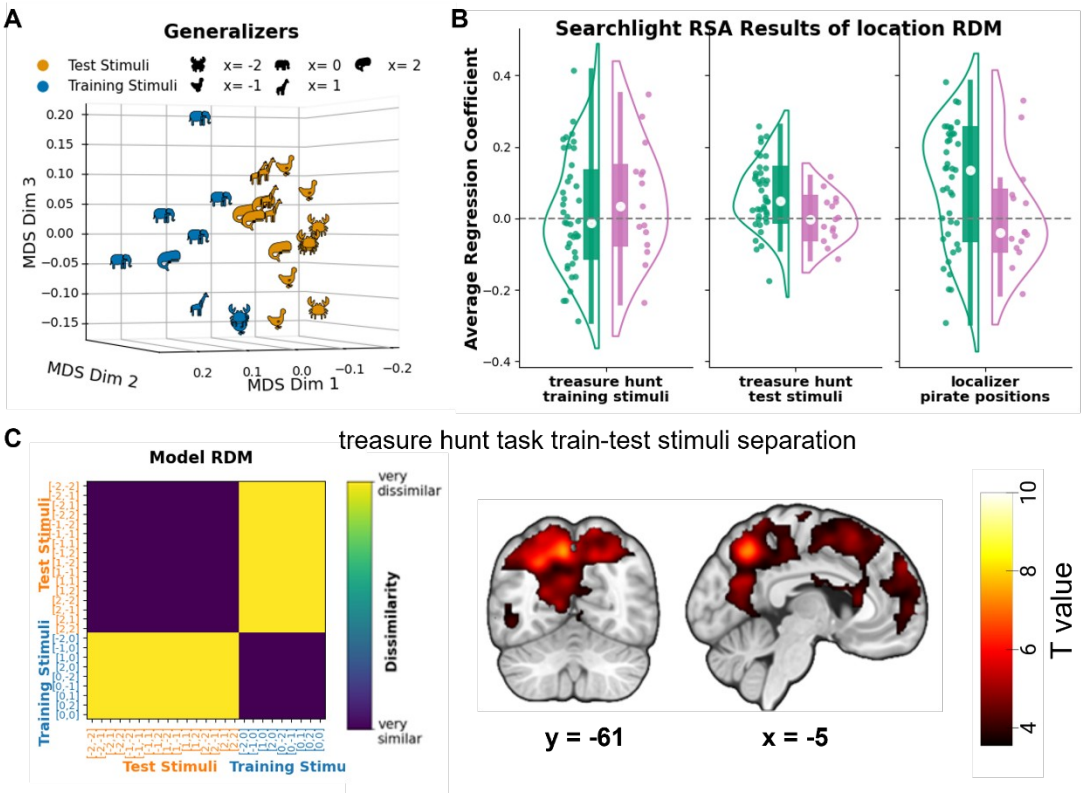
Figure 2-7. Low-dimensional mental map of test stimuli and low-dimensional composition in superior parietal gyrus (SPG). (A) Second-level searchlight result of test stimuli location representation in treasure hunt task for generalizers. Display threshold: cluster-defining threshold $p < 0.001$. The clusters in bilateral superior parietal gyrus were combined into a functional ROI for the analyses in this Figure. (B) Second-level searchlight result of location representation in localizer task for generalizers (Different view of the red cluster from Figure 2-3C). Display



threshold: cluster level $p<0.05$ FDR-corrected; cluster-defining threshold $p<0.0001$. (C) Functional dimensionality of the functional ROI calculated

from activity patterns non-centre training stimuli. (D) Cross-condition feature decoding accuracy. Dots represented the decoder's mean evaluation accuracy across 20 random seed for each participant. (E-F) Predicting test stimulus from training stimuli. (E) prediction quality of retrieval pattern applied to held-out data. (F) Similarity of the empirical retrieval patterns with perceptual matching and distance modulated models. *Notes*: Asterisks indicated results of permutation tests, *p<0.05, **p<0.01, ***p<0.001. In the violin plots (D-F), dots represented individual participants. Box indicated the quantiles in each participant group, while whiskers extended to points within 1.5 IQRs of the lower and upper quartile.

Figure 2-8 Separation of training stimuli and test stimuli in SPG. (A) Multidimensional scaling result from the average SPG neural RDM of generalizers. (B) Visualization of representation similarity with the location RDM for training stimuli, test stimuli from the treasure hunt task and pirate positions in the localizer task. Dots represented the mean regression



coefficient of location RDM in the parietal ROI from searchlight for one participant. Dots represented individual participants. Box indicated the quantiles in each participant group, while whiskers extended to points

within 1.5 IQRs of the lower and upper quartile. To avoid double-dipping, no statistical test at group level was performed for the visualization of average beta weights of location RDM. (C) Second-level searchlight result of train-test separation in treasure hunt task for generalizers. Display threshold: cluster level $p<0.05$ FDR-corrected; cluster-defining threshold $p<0.001$.

2.4.3 Hippocampus represented rules learned from training stimuli using a high-dimensional parallel code

Thus far, we found evidence for low-dimensional representation of test stimuli in regions that were likely to involve in the readout process (V1 for mental imagery or preparatory gaze, precentral/postcentral gyrus for preparatory motor planning and superior parietal gyrus for coordinate transformation, see Table S3-2). The classical regions involved in inference, such as hippocampus and medial prefrontal cortex, were not involved in this low-dimensional spatial code. We examined activity patterns in the anatomically defined ROIs: bilateral hippocampus (HPC, Figure 2-7A) and ventral medial prefrontal cortex (vmPFC, Figure 2-8A). Similar to superior parietal gyrus, HPC and vmPFC represented training and test stimuli in separate manifolds. We were able to reliably decode stimuli condition (training vs test stimuli) from HPC and vmPFC (Figure S2-5, all generalizers were above chance). The RSA searchlight of train-test separation further confirmed that the observed separation was not driven by feature in HPC and vmPFC. Results revealed a cluster in right hippocampus where the peak passed whole-brain correction (peak coordinate [37.5, -11, -14.5], $t(40) = 5.726$), and the vmPFC ROI overlapped with the frontal clusters (left: peak coordinate [-37.5, 46.5, -9.5], $t(40)=7.313$; right: peak coordinate [15,49,0.5], $t(40)=4.72$; see Table S2-1B for a full list of clusters). This motivated us to ask if the rule information was present in these regions and the representation strategy adopted by these regions.

Features of training stimuli can be reliably decoded from HPC representation (Figure 2-7C, all generalizers were above chance for both features). Therefore, we examined the rule representation strategy adopted by HPC. First, we investigated if HPC compressed the rules into two one-dimensional axes in generalizers. In stark contrast to V1 (Figure 2-3C) and superior parietal gyrus (Figure 2-5C), the distribution of functional dimensionality of HPC in generalizers appeared to be more dispersed and was on average higher (Figure 2-7B, paired-sample permutation test on mean dimensionality, HPC vs V1: mean difference=0.695, $p<0.001$; HPC vs superior parietal gyrus: mean difference=0.713, $p<0.001$). One possibility is that HPC data are noisier than V1 and SPG, which could lead to increased dimensionality. To eliminate this possibility, we used the reconstruction quality yielded from the cross-validated SVD for dimensionality estimation as a proxy for signal quality (see 2.3.9 Cross-validated dimensionality estimation). While the data in HPC was noisier compared to V1 and SPG (paired-sample permutation test on SVD reconstruction quality, HPC vs V1: mean difference=-0.124, $p<0.001$; HPC vs superior parietal gyrus: mean difference=-0.143, $p<0.001$), the high dimensionality in our HPC data was not merely due to noise. Instead, it reflected high feature separability in the neural space. After controlling for the signal quality, we observed a positive correlation between functional dimensionality of feature representation and feature decoding accuracy (Figure 2-7C).

Next, we investigated the relationship between the rules. One strategy would be to reflect the orthogonality between the two spatial axes (Figure 2-7D left), i.e. using space as a scaffold. Alternatively, representation reflect the shared relational structure between the rules. This would lead to a parallel representation of axes in which the top of the y-axis would be aligned to the left/right of the x-axis (Figure

2-7D right, example of top-left aligned parallel axes representation). Note that the alignment between axes is completely arbitrary, x- and y-axis could be top-left aligned (leading to a significantly positive PS) or top-right aligned (leading to a significantly negative PS).

To visualize this, we computed the average cosine similarity between coding directions from different axis as the parallelism score (PS) to quantify the alignment between axes for each generalizer. The null distribution was generated by randomly assigning stimulus to an activity pattern, re-calculating the parallelism score and then fitting a Gaussian distribution to the null PS values (10,000 shuffles). We then classified generalizers' alignment scheme according to their PS score (Figure 2-7F). If a generalizer's PS was significantly smaller than the null, they would be classified as parallel top-right aligned. Similarly, if a generalizer's PS was significantly greater than the null, they would be classified as parallel top-left. If a generalizer's PS did not deviate from the null, they would be classified as orthogonal. To visualize the representation of these three groups of generalizers, we visualized the representation of each group of generalizers by performing MDS on the average RDM of each group (Figure 2-7G).

To quantitively test the orthogonal and parallel hypotheses, we ran a cross-validated RSA analysis in HPC ROI. We focused on the neural distance between non-centre training stimuli and compared the neural RDM with three model RDMs: axis separation, orthogonal and parallel. The axis separation RDM was a control RDM that assigned a distance of zero to all within-axis training stimuli pairs and a distance of one to all across-axis training stimuli pairs. Orthogonal RDM described the distance between stimuli as the Euclidean distance on the ground-truth map. Parallel RDM was computed based on alignment scheme (top-left or top-right) and closer distance would be assigned to a pair of stimuli occupying the same axis location (e.g.

x=-2 and y=-2 if top-left aligned). Activity patterns of non-centre training stimuli from different runs were split into fitting and evaluation set. PS was calculated from the average activity pattern on the fitting set to derive alignment scheme for generating the parallel RDM. Then the RSA was performed on the evaluation set. The cross-validated RSA analysis revealed that the neural RDM in hippocampus resembled the parallel RDM (Figure 2-7E). Taken together, our analyses suggest that the hippocampus represented the rules in the training stimuli with two high-dimensional parallel manifolds.

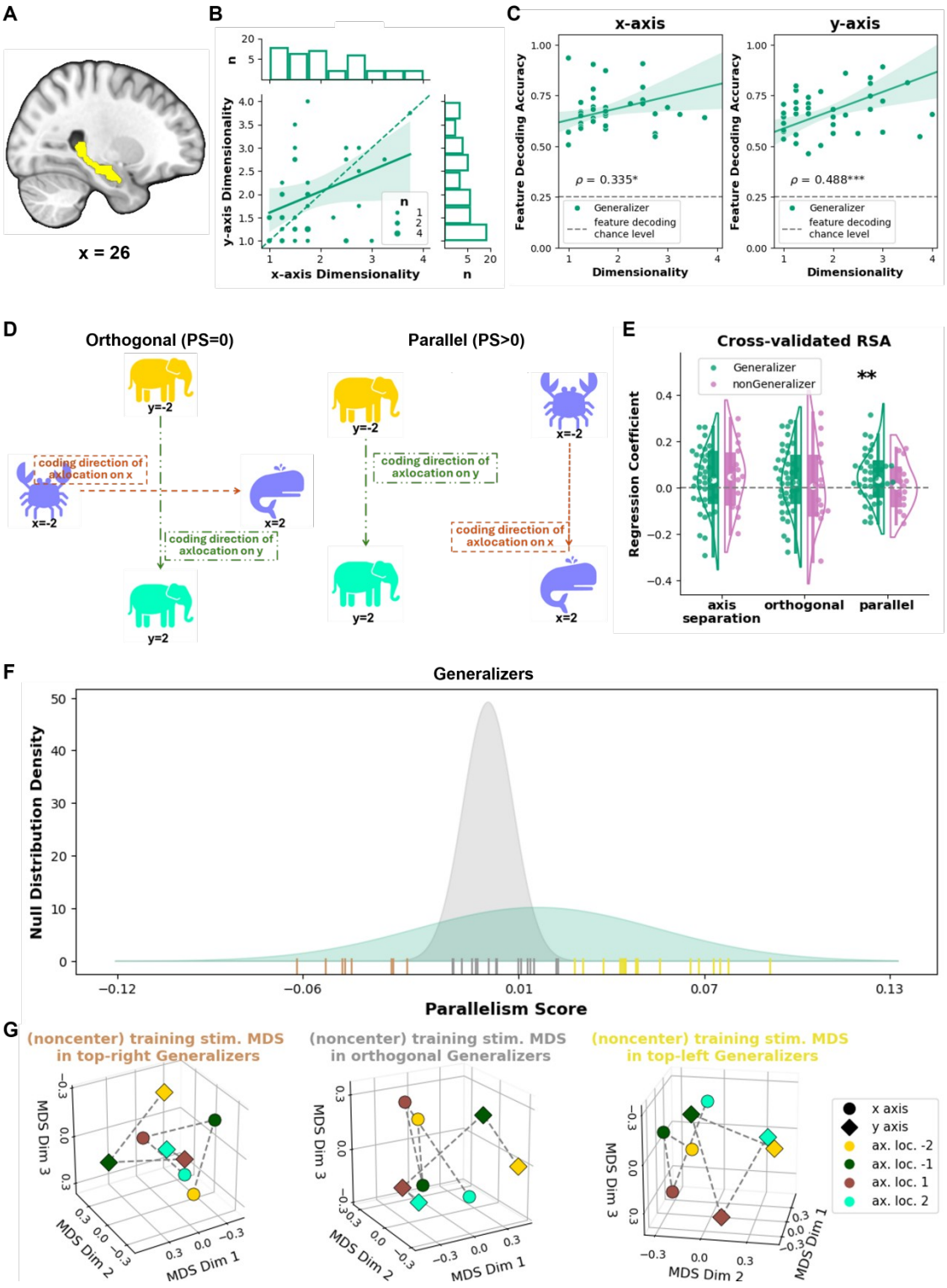


Figure 2-9. High-dimensional parallel representation for the training stimuli in hippocampus. (A) Anatomical mask for hippocampus (B) Functional dimensionality of hippocampus calculated from activity patterns of non-centre training stimuli. (C) Partial correlation between functional dimensionality and feature decoding accuracy after controlling for noise. SVD reconstruction quality was used as a proxy for noise. (D)

Schematic illustration of orthogonal and top-left aligned parallel representation. (E) Cross-validated RSA comparing neural RDM to axis separation, orthogonal and parallel model RDMs in hippocampus ROI. Dots represented individual participants. Box indicated the quantiles in each participant group, while whiskers extended to points within 1.5 IQRs of the lower and upper quartile. Asterisks indicated results of permutation tests, *p<0.05, **p<0.01, ***p<0.001. (F) Distribution of absolute PS in generalizers. Shaded area in grey indicated a fitted Gaussian distribution fitted to the null. Null was generated by randomly shuffling the activity patterns assigned to training stimuli and re-calculating the PS. Coloured sticks indicated the PS of individual generalizers with colours indicating alignment schemes classified according to the relationship between participants' hippocampus PS and the null distribution: parallel top-right, parallel top-left, and orthogonal. (G) MDS of generalizers with different alignment schemes.

2.4.4 vmPFC showed high-dimensional compositional neural code

The previous analysis established that the rules were represented in the hippocampus with high dimensional parallel neural code, but hippocampus didn't put the information together to infer the test stimuli. Retrieval pattern analysis showed that we were not able to predict the test stimulus activity pattern from linear combinations of training stimuli in the hippocampus (Figure S2-4A). Given the abundant connection between hippocampus and vmPFC, we asked if vmPFC acted as the hub for composition.

Apart from the separation in training and test stimuli, another shared representation characteristic between vmPFC and HPC was the high dimensionality for rule representation. The distribution of dimensionality for feature representation in vmPFC was also higher

dimensional than V1 and the superior parietal gyrus (Figure 2-8B, paired-sample permutation test on mean dimensionality, vmPFC vs HPC: mean difference=-0.171, p=0.131; vmPFC vs V1: mean difference=0.524, p<0.001; vmPFC vs superior parietal gyrus: mean difference=0.543, p<0.001). Similarly, the high-dimensionality contributed to better feature separability in the neural space (Figure 2-8C) after controlling for the lower signal quality in vmPFC compared to V1 and SPG (paired-sample permutation test on SVD reconstruction quality, vmPFC vs V1: mean difference=-0.126, p<0.001; vmPFC vs superior parietal gyrus: mean difference=-0.145, p<0.001). Note that the dimensionality for x-axis feature was not correlated with that of the y-axis feature in vmPFC (r=0.216,p=0.192), but we were unsure about the cause of this asymmetry.

However, unlike hippocampus, the cross-validated RSA in the vmPFC ROI didn't reveal a significant correlation between neural RDM in vmPFC and parallel model RDM (Figure 2-8D). Instead, vmPFC ROI representation was significantly correlated with the axis separation and orthogonal RDMs. Therefore, vmPFC represented the task rules in two separated, orthogonal, high-dimensional manifolds.

Moreover, vmPFC showed compositional representation. Feature decoder could also cross-generalize between training and test stimuli, although only for the y-axis feature (Figure 2-8E). Retrieval pattern analysis showed that test stimulus representation could be predicted from linear combinations of training stimuli only in generalizers (generalizers: R^2 = 0.012±0.003, p<0.001; non-generalizers: R^2 = 0.001±0.009, p=0.480) . The retrieval pattern in generalizers resembled the perceptual matching model (β=0.039±0.014, p=0.008) but not the distance modulated model (β=0.019±0.015, p=0.223; Figure 2-8F), suggesting that composition in vmPFC happened in a high-dimensional space.

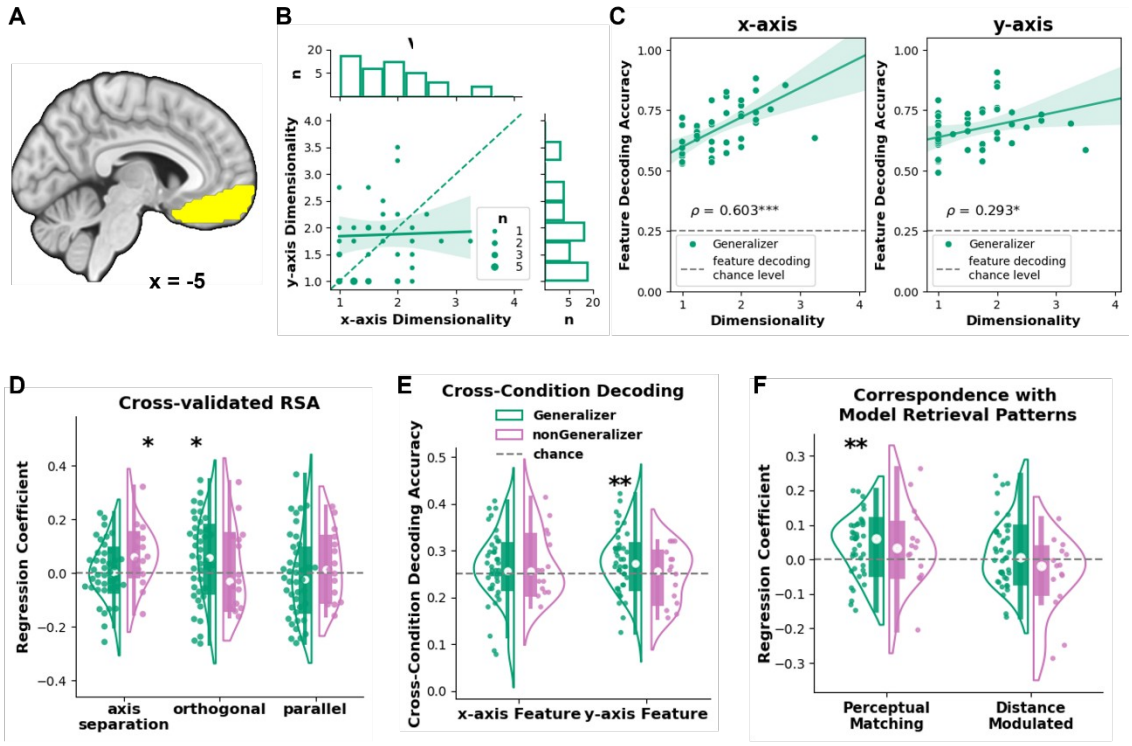


Figure 2-10. High-dimensional compositional representation in vmPFC. (A) Anatomical mask for vmPFC (B) Functional dimensionality of vmPFC calculated from activity patterns of non-centre training stimuli. (C) Correlation between functional dimensionality and feature decoding accuracy. (D) Cross-validated RSA comparing neural RDM to axis separation, orthogonal and parallel model RDMs in vmPFC ROI. (E) Cross-condition feature decoding accuracy. Dots represented the decoder's mean evaluation accuracy across 20 random seed for one participant. (F) Predicting test stimulus from training stimuli. Similarity of the empirical retrieval patterns with perceptual matching and distance modulated models. *Notes*: Asterisks indicated results of permutation tests, *p<0.05, **p<0.01, ***p<0.001. In the violin plots (D-F), dots represented individual

participants. Box indicated the quantiles in each participant group, while whiskers extended to points within 1.5 IQRs of the lower and upper quartile.

2.5 Discussion

In this chapter, we leveraged a compositional rule generalization task to investigate the neural representation of rules and the composition of rules that support successful inference on novel combinations. We found that several regions demonstrated different representation profiles. In the following section, we will synthesize these findings and discuss the implications in terms of the cognitive process involved.

Several regions showed strong separation between the representation of training and test stimuli. In the ROIs we investigated, all participants (generalizers and non-generalizers) showed above-chance decoding accuracy for training vs test stimuli. As strong feature separation could also lead to high train-vs-test decodability, we further ran a whole-brain searchlight RSA where the train-test separation RDM compete with feature RDM and location RDM to explain variance in the neural RDM. While the train-vs-test decodability in visual cortex was indeed driven by feature, it was not the case for the parietal, HPC and vmPFC. While the cause of the separation remains unclear, we speculated that these regions may be involved in the inference process.

We found evidence of hippocampus representing the task rules that reflected the abstract relational structure shared between rules. On a group level, we found that many participants represented rules as two parallel manifolds according to a certain axis alignment scheme (e.g. top of y aligned to left/right of x). Cross-validated RSA also confirmed that the rule representation in hippocampus was better described by a parallel model rather than an orthogonal model. This

is consistent with the role of HPC in maintaining cognitive map of space and abstract relational structure (Donato et al., 2023; O'Keefe & Dostrovsky, 1971; Whittington et al., 2020). The current thesis doesn't have a computation model that can explain why and how this abstraction happened. One future direction would be to incorporate the multiplicative algorithm (tensor product) to implement structural binding across axes (Bakermans et al., 2025).

Hippocampus did not compress the symbolic features into two low dimensional axes. Instead, it maintained a high dimensional code that allowed for feature separation. This echoed with the findings in the previous chapter where we showed that high dimensionality helps maintain feature separability. Agents encoded the input in factorised high dimensional code when simultaneously optimizing for the task objective and an auto-encoder. High dimensionality could also come from a pretraining for feature separability. In the latter case, agents maintain the high dimensional representation of features and learning on the low-dimensional task is driven by plasticity in the readout weight. One possibility is that the observed high dimensionality in the HPC fMRI data could result from the representation prior of participants. In real world, animals and colours reflect distinctive individual experiences and are maintained in separate, independent memories (Gluck & Myers, 1993). The compression of these symbolic features into continuous spatial axes was carried out in the readout process, in our case, it is likely in the parietal cortex.

Although hippocampus showed evidence of rule representation, its representation of test stimuli did not show signature of composition through vector addition. However, this was true for vmPFC. In other aspects, vmPFC had similar profile for rule representation as HPC. It was high-dimensional. Its parallelism score was correlated with that from HPC. But vmPFC demonstrated an additive code for composing

test stimuli from training. Moreover, in line with the high-dimensional rule representation, the vector addition happened in a high-dimensional space, following a perceptual matching model where the relevant training stimuli (those with shared perceptual feature as the test stimulus) contributed more to the composition. This is reminiscent of the additive semantic composition in mPFC from the neurolinguistic literature (Pylkkänen et al., 2014; Pylkkänen, 2019).

So far, we have found evidence of how rules were encoded in a high dimensional parallel space in HPC and composed in vmPFC. Then where is the readout region that compresses and transforms the composed representation into a spatial frame of reference? Using whole-brain searchlight RSA, we found that the superior parietal gyrus, which is the lateral part of the superior parietal cortex (J. Wang et al., 2015) and part of the posterior parietal cortex (PPC, Whitlock, 2017), represented test stimuli as the corresponding treasure locations. It overlapped with the occipital-parietal cluster that represented pirate position in the location localizer task and showed low-dimensional representation of the rules. Moreover, test stimuli representation in this region was compositional in a low-dimensional space. Collectively, our findings showed that PPC adopted a location code for representing the test stimuli. This is perhaps not surprising given PPC's critical role in visuomotor transformation (Buneo et al., 2002). PPC represent space with different reference frames, and serve as a hub for coordinate transformation between these reference frames (Andersen et al., 1985, 1997; Cohen & Andersen, 2002).

Finally, across the whole brain, only two regions represented both training and test stimuli in a low-dimensional spatial map: the precentral gyrus and the visual cortex. Both regions have rich connection with PPC. Together, they could represent the low-dimensional map passed from PPC but with different coordinates

system. The precentral gyrus is part of the motor area that controls voluntary motor movement (Banker & Tadi, 2025). This cluster was not present in the searchlight RSA for pirate position representation. Therefore, the multivariate signal here may reflect the preparation of moving pirate into the treasure location. This finding is reminiscent of the representation of visual stimuli in arm-centric coordinates by premotor cortex in macaque (Graziano et al., 1994). In the visual cortex, the representation was surprisingly low dimensional (while still maintaining feature representation). Moreover, compared to PPC, V1 showed even stronger evidence of a spatial coding scheme shared across task : it is the only region where the location decoder can generalize across tasks. Together, we have strong evidence that at stimuli onset, visual cortex responded as if stimuli were presented in the corresponding treasure location. One interpretation of this finding is that visual cortex forms a mental imagery map. This is in line with the visual working memory literature that showed visual cortex encoded stimuli-response association in working memory tasks both in human and macaque (Christophel et al., 2012; Jonikaitis et al., 2025). Alternatively, it could result from gaze directed towards the future response location. Gaze served as an externalization of information retrieved from memory or composed from inference, or more generally, motor plans (Hommel et al., 2001), encoded in eye-centric coordinates. Unfortunately, we did not record participants' eye-tracking data. Therefore, these two accounts cannot be distinguished based on our current evidence.

Collectively, what the current study revealed could be interpreted as a set of transformation from sensory input to behavioural response to solve the compositional task. The transformation happens in the working memory and is distributed across the brain (Christophel et al., 2017): retrieving rule information from long-term memory in

hippocampus, combining them together in vmPFC, transforming into spatial reference frame in parietal cortex, formation of motor plans in precentral gyrus, and mental imagery or externalization of internal state by oculomotor system in visual cortex.

2.6 Limitations

The evidence on the representation strategy of rules in vmPFC was less conclusive than in HPC. First of all, while dimensionality for both rules not significantly different from HPC, they were not correlated within participants. As the dimensionality was significantly higher than V1 and superior parietal gyrus and the correlation between dimensionality and feature decodability still holds, we nonetheless concluded that vmPFC representation was high dimensional. Secondly, PS of vmPFC was correlated with PS of HPC, but the cross-validated RSA did not reveal a significant correlation between parallel model and vmPFC neural RDM. We refrained from making a conclusion that vmPFC represented the abstract structure between rules in the same way as HPC. It could be that the parallel effect in vmPFC was a byproduct of receiving information from HPC representation. But our current evidence could not verify this speculation.

Based on the representation profiles of different regions and the literature on connections between these regions, we sketched out a potential pathway of how the brain solved the composition rule generalization task. However, we did not have direct evidence for this speculated information flow. While functional connectivity analysis could potentially shed light on this question, it is not straightforward to translate this into the multivariate approach we used for this study.

In terms of task design, the study could be improved in three aspects. First, while we found strong separation between training stimuli and test stimuli in the neural representation space, we did not find a univariate signal for inference, i.e., no regions showed heightened activation during the test stimuli compared to the training stimuli. This may be related to familiarisation with the test stimuli. Participants were already exposed to the test stimuli in the pretraining session. The inference process in generalizers may have become to some extent automated. Future studies could consider only showing a subset of the test stimuli during pretraining to avoid familiarisation.

Secondly, our analyses worked under two assumptions that could be further validated. One was that training stimuli representation largely reflected rule representation. In the analyses on dimensionality and parallelism, training stimuli on different axes were used as the equivalence of the two rules. Due to this, representation of the trained colour and shape could not be dissociated from the representation of rules. In other words, given that blue elephant is the centre stimulus, the representation of x-axis rules cannot be dissociated from the representation of the colour blue. Similarly, the representation of y-axis rules cannot be dissociated from the representation the shape elephant. While it is reasonable to assume that the observed parallelism was not caused by elephant and blue, it would be more robust if future study could obtain representation of the individual component and replicate the current finding. The other assumption we made was that spatial representation would be compositional in low-dimensional space. In the localizer task, we only showed pirates appearing in the training locations. When we composed the test stimuli pattern from these training location patterns, we were essentially assuming that the pattern for location (-

2,-2) could be composed from (-2,0) and (0,-2). But our current design could not validate this presumption.

Finally, a key aspect of learning in this task involves factorising the trained stimuli-location associations into two independent rules (regardless of whether they are represented orthogonally or in parallel). However, the stimuli were composed from colour and shape, which are already internally factorised into distinct representational dimensions by the perceptual system. This made it difficult to disentangle the factorisation induced by the successful learning and the factorisation induced by perception. In the current study, we did not focus on the factorisation aspect. Future studies could use other features to get around this problem. In the next chapter, we will introduce a revised task design in the hope of addressing some of these limitations.

Chapter 3 Solving compositional rule generalization task with low and high dimensional representation

Contents

3.1 Introduction.....

3.2 Experiment 1: Inducing high dimensional representation with auxiliary feature reconstruction loss.....

3.2.1 Methods.....

3.2.2 Results.....

3.2.3 Summary.....

3.3 Experiment 2: Inducing high dimensional representation with unsupervised pretraining.....

3.3.1 Methods.....

3.3.2 Results.....100

3.3.3 Summary.....101

3.4 Experiment 3: The flexibility of high-dimensional representation.....101

3.4.1 Method.....101

3.4.2 Results.....103

3.4.3 Summary.....105

3.5 Summary & Discussion.....105

3.1 Introduction

In everyday life, humans often receive high-dimensional inputs such as visual scenes or sentences. However, the actions required are often low-dimensional, e.g. making a yes/no decision or determining where to go. Identifying the underlying statistical structure in high-

dimensional data for efficient readout is a core aspect of learning. A natural solution is to perform dimensionality reduction, i.e., to discard irrelevant features for the readout and extract the principal components of the data. This allows for more robust generalization performance and is protective against task-irrelevant input changes and noise. In object recognition, this is thought to be the basis of perceptual invariance (DiCarlo & Cox, 2007; W. Wang & Carreira-Perpinan, 2014). In continual learning, dimensionality reduction allows for abstract representation of shared structures across different tasks that can be flexibly re-used (Johnston & Fusi, 2023; Zhou et al., 2021). However, in the previous chapter, we found that hippocampus and vmPFC did not compressed the colour and shape representations to two one-dimensional axis corresponding to the spatial readout. This is in line with previous findings that the prefrontal cortex and hippocampus have been found to maintain abstract yet still high-dimensional representation (Bernardi et al., 2020; Tang et al., 2019). One theoretical framework suggests that such high-dimensional representation may arise from mixed selectivity of neurons and is more flexible as it allows for more possible read-outs by downstream regions (Fusi et al., 2016).

In this simulation study, we considered learning as an optimization process (Hennig et al., 2021). Instead of thinking about high-dimensional representation as a result of biological constraints (e.g. mixed selectivity), we asked what would be a potential natural drive in the optimization process that may lead to a high-dimensional solution. We trained MLPs to perform the treasure hunt task which required mapping from high-dimensional features to low-dimensional spatial readout (Dekker et al., 2022). First, we conducted two simulation experiments to show that 1) the task can be solved with both low- and high-dimensional representation, and 2) high-

dimensional representation may come from the need to maintain input separability. Then, we demonstrated the flexibility of high-dimensional representation in faster learning of new input-output mapping rules. More importantly, such high dimensionality and high input separability were maintained during learning of low-dimensional tasks. Faster learning was facilitated by developing a new readout rule instead of slowly manipulating how input was being encoded.

3.2 Experiment 1: Inducing high dimensional representation with auxiliary feature reconstruction loss

3.2.1 Methods

Task set up

The treasure hunt task was setup in the same way as in Chapter 2. In this task, agents were provided with cues and were asked to predict treasure locations based on the cues. Cues were coloured animals. A total of 25 stimuli were formed by the factorial combinations of five shapes and five colours. The hidden rule of the task was that each feature of the cue (colour or shape) indicated location on either the x- or the y-axis in cartesian coordinates (Figure 3-1A). In essence, the task required mapping from high-dimensional discrete features to a two-dimensional (2D) continuous space. Neural network models were trained on the 9 training stimuli on the central “+” locations and tested without feedback on the remaining 16 stimuli.

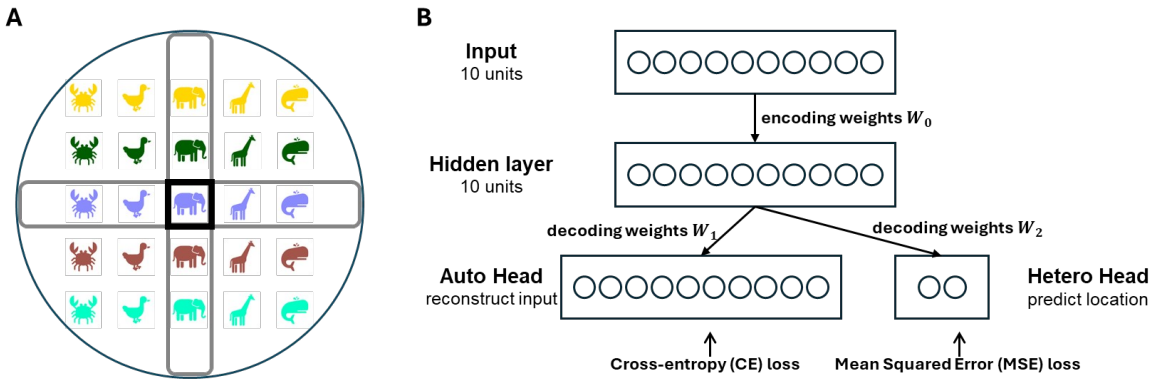


Figure 3-11.Task set up and neural network agent. (A) The map of the treasure hunt task. In this example, shape is mapped onto x location and colour is mapped onto y location. The stimuli lying on the central "+" surrounded by the grey box are the training stimuli. The rest of the stimuli are test. (B) Network setup

Neural network simulation

We built neural network agents to perform the treasure hunt task using PyTorch. The agents' architecture was a feedforward multi-layer perceptron (MLP) with one hidden layer (Figure 3-1B).

The input layer had 10 units corresponding to the 10 features (five shapes and five colours) that were used to compose the stimuli. If a feature was present in the combination, the corresponding input unit in the input vector x would be set to 1. The input layer was fully connected to the hidden layer with the encoding weights W_0. The hidden layer had 10 units, and a linear activation function was used for the hidden layer units without bias terms.

The hidden layer was then fully connected to the output layer. The output layer consisted of two heads: the hetero-encoder head and the auto-encoder head. The hetero-encoder head had two units and was trained to predict the location of treasure given the feature combination. The additional auto-encoder head had the same number of units as the input layer. It was trained to reconstruct the combination of features from the input. Hence there were 12 output

units in total. The behaviour of this artificial agent was fully deterministic. The auto-encoder head made responses using the argmax rule, i.e. the colour that corresponded to the unit with the largest y across all five colour units was chosen as the response for colour, same for shape.

Neural network models were trained using stochastic gradient descent and a combined loss. In this combined loss, the hetero-encoder was trained on the mean squared error (MSE) loss whereas the auto-encoder was trained on cross-entropy (CE) loss. The MSE loss was calculated as:

$$MSE\,loss(y,t) = -\sum_{i=1}^{2} (t_i - y_i)^2$$

t was the target vector (supervision signal) for the hetero-encoder, y was the output vector of the hetero-encoder with y_i being the output value of the i^{th} output unit. For CE loss, the implementation in PyTorch was equivalent to the combination of log-SoftMax and negative log-likelihood loss.

$$CE\,loss(y,t) = -\sum_{i=1}^{12} \log \frac{\exp(y_i)}{\sum_{j=1}^{12} \exp(y_j)} * t_i$$

t was the target vector (supervision signal) for the auto-encoder, y was the output vector of the auto-encoder with y_i being the output value of the i^{th} output unit.

Importantly, we manipulated the weight assigned to the CE and MSE loss to see if we could change the hidden layer representation dimensionality.

$$Loss = \alpha * CE + (1-\alpha)\,MSE$$

The parameter α indicated the weight assigned to autoencoder training. When $\alpha = 0$, the agent was a pure hetero-encoder only trained

to perform the treasure hunt task. When $\alpha=1$, the agent was a pure autoencoder only trained to reconstruct the input. When $0<\alpha<1$, the agent was trained with mixed objectives and should be good at both location prediction and feature reconstruction. Following early theoretical work on MLP as feature extractors (Bourlard & Kamp, 1988; Ghahramani, 1996; Kung et al., 1994), we can derive the theoretical value of hidden layer dimensionality from the cross-covariance matrix of input and output given different values of α. We can expect the following at the time of convergence to the global optimum:

1) For the pure hetero-encoder agent ($\alpha=0$) the dimensionality should be two.
2) For the pure auto-encoder agent ($\alpha=1$), the dimensionality should be eight.
3) For the mixed-objective agents ($0<\alpha<1$), the dimensionality should be eight.

Each epoch started with 9 training trials (1 trial per training stimulus). After that, generalization performance was evaluated using the remaining 16 test stimuli without feedback. Supervision signals were only provided during training trials. Weights were updated after each training trial.

We ran simulations with varying auto-encoder loss weight (α uniformly sampled from 0 to 1 at a step size of 0.05). 30 runs were collected for each auto-encoder loss weight α. Weights were initialised randomly in each run. Encoding weights (W_0) were initialised in the rich regime (Flesch et al., 2022), sampled from Gaussian distributions with means of 0 and variances of 0.01. Readout weights (W_1 and W_2) were initialised with variances of $1/\sqrt{n_h}$, n_h being hidden layer size. The same random initialization was used across the alpha configurations in the same run.

Representation similarity analysis

To determine the information encoded by the agent, we performed representation similarity analysis. First, we obtained the activity pattern matrix of an agent by taking the hidden layer activation resulting from all the stimuli. Then, we computed the dissimilarity between a pair of stimuli using correlational distance. This procedure yielded a 25×25 representation dissimilarity matrix (RDM).

We also derived two model RDM, cartesian and feature based on the information that could be encoded by the agent, i.e., spatial location and features. In the cartesian model RDM, the distance between a pair of stimuli corresponded to the Euclidean distance between their truth treasure locations. In the feature model RDM, the distance between a pair of stimuli inversely correlated to their number of shared features. Then we regressed the agent's RDM on these two model RDM to determine which model RDM best fit the hidden layer representation. Note that only the lower triangular part of the matrices was used. Model RDMs were standardized before running regression.

3.2.2 Results

Hetero-encoder agents and mixed objectives agents showed indistinguishable generalization performance in location prediction

After training, agents with non-zero task objective weights ($\alpha<1$) learned to predict location based on the training stimuli and generalize zero-shot to the test stimuli (Figure 3-2A-B). In the meantime, mixed objectives agents ($0<\alpha<1$) were able to reconstruct the input to the same extent as the pure auto-encoder agent ($\alpha=1$) while the pure hetero-encoder agent failed to do so (Figure 3-2C).

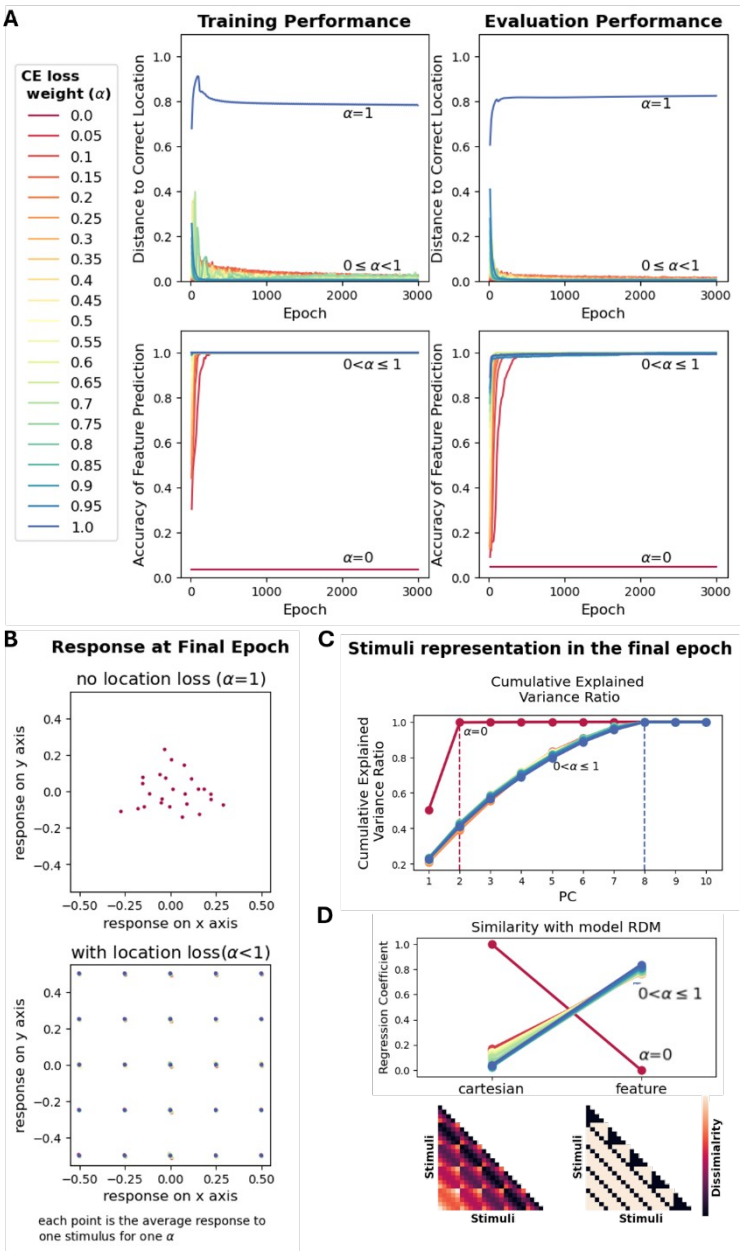


Figure 3-12. Behavioural performance and hidden layer representation. (A) Location prediction performance measured by distance to correct location on each axis at the final epoch for each network.

Auxiliary feature loss forces NN to form a high-dimensional representation

Next, we examined the hidden layer representation of training stimuli at the final epoch. First, we performed PCA to see how many principal components were required to explain the variance in hidden layer activity (Figure 3-2C). As predicted in the methods section, the pure

hetero-encoder agent formed a two-dimensional representation to solve the task. Meanwhile, the mixed objective agents ($0<\alpha<1$) formed a high-dimensional solution similar to the pure auto-encoder agent ($\alpha=1$). This result was further confirmed by a representation similarity analysis (Figure 3-2D). We computed the representation dissimilarity matrix (RDM) of an agent based on the pairwise correlation distance on the hidden layer activation vectors resulting from all stimuli. Next, we regressed the agent's RDM on two model RDMs: cartesian and feature. The cartesian RDM depicted a low-dimensional solution that represented stimuli as cartesian coordinates and the RDM reflected the Euclidean distance between stimuli. The feature RDM depicted an uncompressed high-dimensional solution that represented stimuli as feature combinations and the RDM reflected the number of shared features. The pure hetero-encoder ($\alpha=0$) represented stimuli as their cartesian coordinates whereas the mixed objective agents ($0<\alpha<1$) represented stimuli as feature combinations in a similar way as the pure auto-encoder ($\alpha=1$).

3.2.3 Summary

In summary, we showed that the need to maintain input separability can elicit high-dimensional representation in agents trained to solve a low-dimensional task. This was implemented by adding auxiliary feature reconstruction loss during the learning process. Agents trained with this auxiliary loss (mixed-objective agents) demonstrated similar successful generalization behaviour as the pure-hetero agent trained without this auxiliary loss.

3.3 Experiment 2: Inducing high dimensional representation with unsupervised pretraining

The feature reconstruction loss in Experiment 1 mimicked the scenario in which agents learn to map unfamiliar inputs to low-dimensional output space while simultaneously acquiring input separability. However, in many cases, agents learn to map familiar inputs to novel outputs. For example, in the treasure hunt task used in Chapter 2, participants learned to map colours and shapes onto treasure locations. In such cases, familiar inputs would already be represented distinctively in a high-dimensional space. Experiment 2 further seeks to study how learning process shapes prior representations in such scenarios.

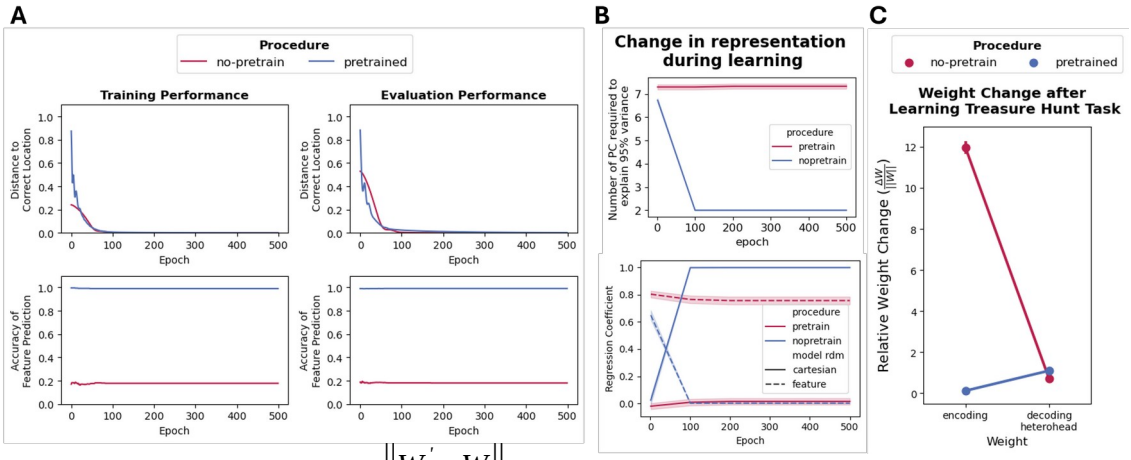
3.3.1 Methods

NN simulation

The architecture was the same as in Experiment 1. However, we did not manipulate the weight assigned to the CE loss of the auto-encoder head. Instead, we pretrained the auto-encoder head to distinguish among the 10 input features. After this initial pretraining stage, we froze the decoding weight of the auto-encoder head and trained the hetero-encoder head and encoding weight on the treasure hunt task. In the control procedure, we did not include any pretraining on features, NN is directly trained on the treasure hunt task. We collected 30 runs with different random initializations. For each random initialization, we collected data from both the pretrain and no-pretrain procedures

Quantifying Weight Change

For each agent, we computed the relative weight change (RWC) after learning the treasure hunt task. We first obtained the encoding weight and the decoding weight of hetero-head before and after task learning. Let W' and W be the weight after and before task learning,



RWC was computed as $\frac{\|W^{'}-W\|}{\|W\|}$.

Figure 3-13. Neural network performance and representation geometry during treasure hunt task learning in Experiment 2. (A) Neural network's performance during training and evaluation during task learning. The top panels show the error in treasure location prediction. The bottom panels show the accuracy of reconstructing feature combinations from inputs. (B) Change in hidden layer representation during task learning. The top panel show the number of principal components required to explain 95% of the variance in the hidden layer activations. The bottom panel show the representation similarity analysis throughout training. The shaded area indicates standard error. (C) Weight change in different layers as a result of treasure hunt task learning. Error bars indicate standard error.

3.3.2 Results

Learning and Generalization Behaviour

In both procedures, the agents learned to predict location based on the training stimuli and generalize zero-shot to the test stimuli (Figure

3-3A top panels). Notably, the pretrained agents maintained their ability to reconstruct features from inputs even though the auto-head was not trained during task learning (Figure 3-3A bottom panels).

High-dimensional representation induced by pretraining was maintained during task learning

We analysed the change in representation during treasure hunt task learning. We calculated the number of principal components required to explain 99% of the variance in the hidden layer activation at different epochs of task learning (Figure 3-3B top panel). In addition, we performed the same representation analysis as in Experiment 1 and showed the change in similarities between the agents' representation and different model RDMs (Figure 3-3B bottom panel).

The pretrained agents and the unpretrained agents demonstrated different representation changes in response to task learning. First, feature pretraining induced high-dimensional representations that kept features distinct from each other in a similar way as the mixed-objective agents in Experiment 1. More importantly, this high-dimensional feature representation was maintained even during task learning, echoing the behavioural result that the pretrained agents maintained their ability to distinguish among inputs (Figure 3-3B, blue line). In other words, learning the 2D location task did not compress the hidden representation in the pretrained agents. Instead, task learning was realised by the change in the readout weight of the hetero-head while the encoding weight remained stable (Figure 3-3C, blue line). Contrarily, in agents that were not pretrained, task learning was mainly driven by a change in the encoding weight (Figure 3-3C, red line). This resulted in compression in representation dimensionality and led to a 2D spatial representation (Figure 3-3B, red line) similar to the pure hetero-encoder agent in Experiment 1.

3.3.3 Summary

In summary, agents undergoing both procedures successfully solved the treasure hunt task, but with different representations. Pretraining the auto-encoder head made agents form high-dimensional feature representations similar to the mixed-objective agents in Experiment 1. Later on, during treasure hunt task learning, these agents developed new readout weights without changing the encoding weights. This allowed the feature representation to be maintained. In comparison, networks that were not pretrained behaved in a similar way as the pure hetero-encoder agent in Experiment 1. They learned the task by compressing the representation to reflect the spatial location of stimuli in 2D space

3.4 Experiment 3: The flexibility of high-dimensional representation

In Experiments 1-2, we established that agents could solve the task equally well with high-dimensional and low-dimensional representation. Next, we examined the benefit of maintaining high dimensional representation.

3.4.1 Method

Transfer Experiment

We designed two types of transfer conditions, reversal and derangement (Figure 3-4A), to test the transfer performance of the trained agent. In the reversal condition, we flipped the sign of feature-to-location mapping. For example, the original mapping from colours to y locations (from top to bottom) was blue-brown-purple-green-yellow, and after reversal, this became yellow-green-purple-brown-

blue. Reversal would not change the mapping at the central location. In the derangement condition, we permuted mapping in a way that no element appears in its original position. We also kept the centre locations unchanged in the derangement condition to match the reversal condition.

We adopted a within-group design in which one of the axes would be reversed while the other would be deranged. We hypothesized that learning of the reversal axis should benefit from shared rule structure and it should be faster compared to the derangement axis.

NN Simulation

We took the trained agents from Experiment 1 and trained them on the transfer experiments. During the transfer learning stage, only the hetero-encoder head was trained, whereas the decoding weight of the auto-encoder head was frozen.

Human Experiment

We collected data from 76 participants on Prolific. The human experiment setup was the same as the simulations. Participants only received feedback on the correct treasure location for the 9 training stimuli on the central “+” locations and were tested on the remaining 16 stimuli without feedback (Figure 3-1A, see Figure 3-1C for trial timeline). First, participants were trained on the original map for 6 train-test cycles. In each cycle, they received two blocks of training on the training stimuli with feedback, and one block of test on the held-out stimuli without feedback. Then, the experiment progressed to the transfer map without explicit cues. In the transfer stage, participants underwent 8 train-test cycles. In each cycle, they received one block of training with feedback on the central “+” locations stimuli and one block of test without feedback.

Participants were classified into learners and generalizers according to their performance in the final cycle of the original map following the same procedure as Dekker et al. (2022). Non-learners were excluded from the following analyses. This yielded 24 generalizers and 23 non-generalizers in total.

Measuring transfer learning rate

To quantify this, we fitted the exponential decay function ($error_c = N_0 \times e^{-\lambda \times c}$) to the agent's error curve for each axis. Here, $error_c$ indicated the error on a given axis in cycle c, and there were two free parameters: the initial error (N_0) and the decay rate (λ).

3.4.2 Results

Faster Transfer Learning in the Reversal Condition in Human and NN Generalizers

In human participants, both generalizers and non-generalizers of the original map learned the training stimuli on the transfer map, but only the generalizers were able to generalize to the test stimuli (Table 3-1, Figure 3-4B). All Neural networks learned and generalized in the transfer map (Figure 3-4C). More importantly, human generalizers as well as NN generalizers of the original map (the pure hetero-encoder and mixed objective agents) showed differential learning dynamics for reversal and derangement conditions. In both human and NN generalizers, learning on the reversal axis reached asymptote earlier than the derangement axis. In contrast, human and NN non-generalizers (the pure-hetero agent) did not show such a difference.

Table 3-1. Linear mixed effect model on error during the transfer learning in human participants.

Variable	Training Stimuli Error (coef.±s.e.)	Test Stimuli Error (coef.±s.e.)
Participant (G>nG)	-0.097±0.033 **	-0.243±0.046 ***
Cycle	-0.047±0.004 ***	-0.017±0.006 **
Participant×cycle	-0.005±0.005	-0.037±0.008 ***

G: generalizers; nG: non-generalizers.
***: p<0.001; **: p<0.01; *: p<0.05.

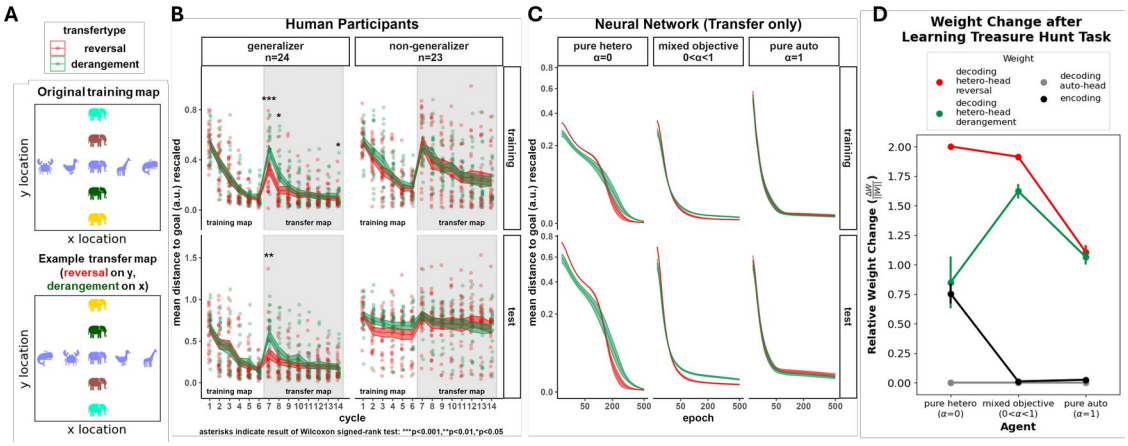


Figure 3-14. Human and neural network performance in Experiment 3. (A) Example maps of training stimuli for the original training map stage (top) and transfer map stage (bottom). In this example, during the transfer stage, the mapping from colour features to y-axis locations was reversed. Meanwhile, the mapping from shape features to the x-axis location was randomly permuted (derangement). (B) Human participants' performance on the original map and the transfer map. (C) Experiment 1's NN performance on the transfer map. (D) Weight change in different NN after transfer learning. The shaded area and error bars indicate standard error.

Faster Transfer Learning in High-Dimensional Neural Networks

Recall from Experiment 1 that the mixed objective agents were generalizers with high-dimensional representation and the pure hetero-encoder agent was low-dimensional. During the transfer stage,

the mixed objective agents reached their asymptote earlier than the pure hetero-encoder agent (Figure 3-4C). To quantify this, we fitted the exponential decay function ($error_c = N_0 \times e^{-\lambda \times c}$) to the agent's error curve for each axis. Here, $error_c$ indicated the error on a given axis in cycle c, and there were two free parameters: the initial error (N_0) and the decay rate (λ). Mixed effect models revealed that mixed-objective agents had a significantly higher decay rate than the pure hetero-encoder agent (Table 3-2).

The advantage of high-dimensional networks during transfer can be explained by the change in weight (Figure 3-4D). High-dimensional agents only changed their readout weight during transfer learning. On the contrary, transfer learning in the low-dimensional agent was largely driven by the change in encoding weights.

Table 3-2. Error decay rate during the transfer learning in pure hetero-encoder and mixed objective networks.

Variable	Training Stimuli Error (coef. ±s.e.)	Test Stimuli Error (coef. ±s.e.)
Transfer (R>D)	0.006±0.011	0.003±0.011
Agent (M>H)	0.029±0.047 ***	0.190±0.026 ***
Transfer× Agent	0.027±0.011 *	0.007±0.011

R: reversal; D: derangement.
M: mixed-objective agents; H: pure hetero-encoder agent
***: p<0.001; **: p<0.01; *: p<0.05.

3.4.3 Summary

In summary, both agents and human generalizers who learned the rule on the original map learned the reversal condition faster than the derangement condition, suggesting a re-use of the previous structure. Comparison between the learning dynamics of the mixed objective agent and the pure hetero-encoder revealed the benefit of high-dimensional representation in flexibly remapping features to output according to new rules.

3.5 Summary & Discussion

In this study, we investigated the potential drive during the learning process to form a high-dimensional solution even though the required readout is low-dimensional. We showed that the need to maintain input separability can elicit high-dimensional representation in artificial agents trained to solve a 2D task. Furthermore, high-dimensional agents learned the 2D task by developing new readout weights without compressing their representation. Finally, high-dimensional agents were more flexible in remapping learned features to outputs according to new rules.

Our results can be linked to several ideas in the representational learning literature.

First, these results confirmed the prediction from the theoretical framework that suggests high-dimensional representation allows for more possible readout by downstream regions (Fusi et al., 2016). We demonstrated that in high-dimensional networks, new learning can be facilitated by merely changing in readout weight instead of changing both encoding and readout. During learning, the representation system often needs to strike a balance between compressing irrelevant features for the current task demand without being overfit

to the current task. The auxiliary loss we embedded in the network help maintain such a balance.

Unsupervised pretraining is commonly used as a technique to facilitate learning efficient and generalizable representations. By leveraging structure inherent in raw input data without requiring explicit labels, unsupervised learning enables models to uncover latent feature dimensions of the input data (Olshausen & Field, 1996; Hinton & Salakhutdinov, 2006). These representations can then serve as a foundation for faster and more efficient supervised learning, which often requires readout based on the latent feature dimensions. This is particularly helpful when supervised examples are scarce (Erhan et al., 2009, 2010). In developmental contexts, integrating unsupervised exposure with sparse supervised labels has been shown to support category learning in infants (LaTourrette & Waxman, 2022). Similarly, recent work in adults suggests that when unsupervised pretraining that aligns participants internal representation axis with experimenter-defined task structure enhance category learning (Bröker et al., 2022). Relatedly, recent work in mice also demonstrated that representation change in visual cortex during task learning is largely driven by unsupervised exposure to naturalistic stimuli and that such unsupervised exposure supports more rapid and effective supervised learning later on (Zhong et al., 2024).

Finally, our study is also related to the line of work which concerns the role of high-dimensional representation in context-dependent behaviour. High dimensional representation induced by mixed selectivity which allows for simultaneous encoding of stimuli and context has been suggested to subserve fast switching between different sensory-motor mapping rules. This speaks to a solution for continual learning in which stimuli in different contexts/tasks are represented in orthogonal subspaces to prevent interference (Flesch

et al., 2022; Kim et al., 2025; Salinas, 2004). However, our simulation did not adopt a contextual cue for rule-switch. It points to a different solution in which contextual separation happened at readout instead of encoding. An interesting future direction would be to see how these different solutions can be mapped onto the biological brain. Indeed, recent evidence suggests that the auditory cortex in mice did not show sensory sharpening after task learning. Instead, a subpopulation of auditory cortex neurons gradually developed conjoint coding of stimulus and reward prediction signals showcasing the emergence of stimulus-action knowledge (Drieu et al., 2024). Recent studies involving rule switching and continual learning in mice demonstrated separate representations of readout rules in different task contexts in anterolateral motor cortex (ALM). More specifically, they showed that ALM created different preparatory states for the same set of stimuli in different contexts (Y.-T. Chang et al., 2024; Kim et al., 2025). These findings support the idea that new learning can be achieved by forming new readout mappings from a stable representational basis, aligning with the mechanism demonstrated in our model.

Chapter 4 Flexible cross-context compositional generalization

Contents

4.1. Introduction....108
4.2. General Methods....112
4.2.1 Procedure....112
4.2.1 Stimuli generation and stimuli subsets....113
4.2.2 Task Pool....115
4.2.3 Performance Metrics....118
4.3. Pilot M1V1: one map task with the symbol combination stimuli....118
4.3.1 Methods....118
4.3.2 Results and Discussions....123
4.4. Pilot M2V1: multi-map task, blocked evaluation....124
4.4.1 Methods....124
4.4.2 Results and Discussions....127
4.5 Pilot M2V2 and pilot M2V3: multi-map task, randomised evaluation, new classification method....130
4.5.1 Methods....131
4.5.2 Results....139
4.6. Symbol representation in the cluster task....143
4.7. Summary and Discussion....145

4.1. Introduction

In this chapter, we set out to discuss a series of behavioural experiments that gradually refine the design of the compositional rule

generalization task in the previous chapter to address its limitations and to explore further questions.

1) Lack of evidence of learning induced factorisation

An important aspect of learning is to disentangle the different independent factors that generate the observed data in the world. This is known as factorisation in neural representations. In the context of our task, successful generalization requires identifying the correspondence between sensory dimensions and spatial axes. In the previous chapter, colour and shape were used. However, they are inherently different during sensory processing. Neurophysiology and neuroimaging studies in macaques and humans have shown that the pathways involved in colour and shape perception are separated at different stages, and their processing is often more interdigitated at the early stage of visual processing and becomes more anatomically segregated at later stages (Livingstone & Hubel, 1988; Sincich & Horton, 2005; K. Seymour et al., 2010; Lafer-Sousa & Conway, 2013). In other words, colour and shape dimensions are internally factorised. Therefore, it becomes difficult to disentangle the factorisation caused by perceptual processes and the factorisation induced by learning in the previous fMRI study.

To address this limitation, in this chapter, we switch from the internally factorised colour and shape dimensions to a single feature dimension (i.e. shape). Various symbol shapes are chosen and further divided into two sets. One set is mapped onto the x-axis locations, and the other set mapped onto the y-axis locations. Stimuli are generated from the factorial combination of these two sets of shapes.

The rationale for generating stimuli from a single feature dimension is to ensure that, prior to learning, all symbols are represented within the same high-dimensional space. Successful generalization requires

splitting the symbols into two sets that corresponded to the two spatial axes. Correspondingly, in the representational space, we should only observe a separation between these sets after learning. In other words, the factorisation should only be observed after learning.

2) Flexible cross-context generalization

In the previous fMRI study, participants learned about the rules on the same map simultaneously, making interpolation of novel combinations was relatively straight-forward. To introduce a more challenging condition, we now require participants to flexibly combine rules learned at different times to make inferences. To test this capacity for flexible, cross-context generalization, we incorporate multiple maps into the experimental design (Figure 4-1). Four sets of symbols are taken, with two mapped onto the x-axes (X1, X2) and the remaining two mapped onto the y-axes (Y1, Y2). X1 and X2 each has five symbols corresponding to locations [-2,-1,0,1,2] on the x-axis. Similarly, Y1 and Y2 each has five symbols corresponding to locations [-2,-1,0,1,2] on the y-axis. These sets form two training maps: one from X1Y1 combinations (training map 1, referred to as *T1* hereafter) and the other from X2Y2 combinations (training map 2, referred to as *T2* hereafter). Participants would be trained and evaluated on these training maps as in the previous fMRI experiment. After training, cross-map generalization tests are introduced to assess whether participants can combine rules learned in different map contexts. Specifically, participants are asked to infer the treasure locations of stimuli composed of symbol pairs drawn from different training maps. Two cross maps can be generated: stimuli in the first cross map (C1) are generated from X1Y2 combinations and stimuli in the second cross map (C2) are generated from X2Y1 combinations.

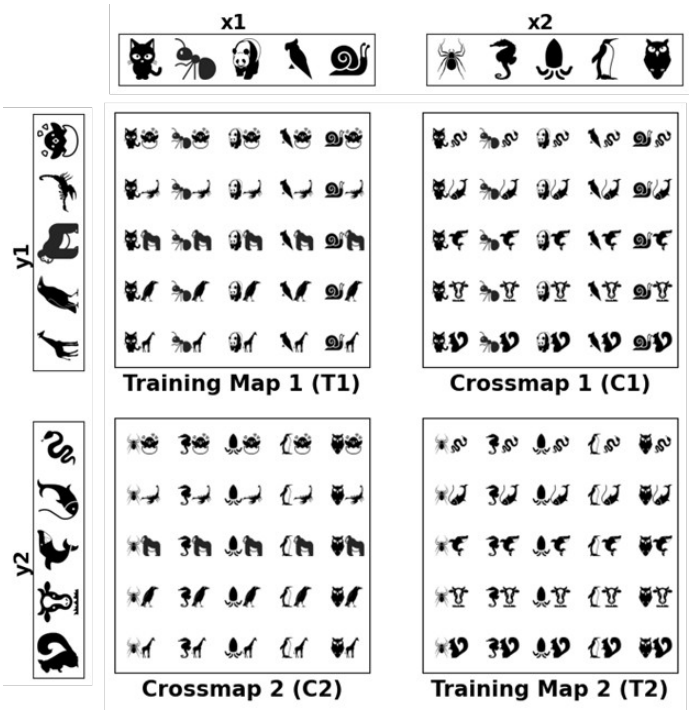


Figure 4-15. Example multi-map stimuli in their ground-truth locations.

Under this multi-map design, the number of stimuli grows to 100. The slow timeline of treasure hunt task (>10s per trial) makes it difficult to have many repetitions for each stimulus. Therefore, it is no longer suitable for a fast event-related design for the increased stimuli set. To alleviate this problem, we designed a fast-paced two-alternative forced choice (2AFC) task where participants were given a limited amount of time (1s) to choose the correct location of the treasure.

Finally, to avoid over-familiarisation with the inference trials involving non-training stimuli, we decided to only present participants with half of these stimuli during pre-training (referred to as validation stimuli hereafter). The other half would only be presented when participants performed the scanner task (referred to as test stimuli hereafter). Another benefit of splitting non-training stimuli is to allow for cross-validation in participant classification procedure. In all our pilot experiments, we will classify participants as generalizers or non-

generalizers based on whether they can successfully infer the treasure location of validation stimuli or not.

3) cleaner test of rule representation

In the previous chapter, we equated rule representation with the representation of training stimuli. Therefore, the representation we obtained would be the conjunction of two features. In the new design, all stimuli are generated from symbol combinations, so it becomes possible to design an additional scanner task to measure the representation of individual components (i.e., shape symbols). This would allow for a clean test of rule representation independent of the task that required composition.

4) Lack of behavioural evidence of parallel representation

In the previous chapter, we found that a significant portion of generalizers represented the rules in parallel to each other in the hippocampus, while the others represented the rules orthogonally. But we did not find behavioural markers for parallel versus orthogonal representation. We hope to design a new behavioural task that can achieve this goal.

In summary, we modify the compositional rule generalization task in several aspects, including generation of new stimuli using symbol combinations, extension to a multi-map setting, and introduction of a two-alternative forced choice (2AFC) task more suitable for an fMRI study. We set out to test the validity of this new design. Moreover, we evaluate whether the 2AFC task can be used to measure generalization performance, i.e. whether performance in the 2AFC task is consistent with that of the treasure hunt task. Finally, we explore whether a behavioural marker for parallel/orthogonal representation can be identified.

4.2. General Methods

4.2.1 Procedure

We collected data for all the pilot experiments online from Prolific. The behavioural experiments were approved by the Medical Sciences Research Ethics Committee of the University of Oxford (R50750/RE009).

All the pilot experiments had similar procedures but varied slightly in the number and order of blocks. We outlined the general procedure below. The specific details of each experiment will be described in the corresponding section.

In general, all experiments consisted of two days (Figure 4-2E). Day 1 was a pre-training stage using the treasure hunt task. The pre-training stage consisted of multiple training-validation cycles. Each cycle consisted of two training blocks followed by one validation block. In the training blocks, participants learned the rules with supervision on the training stimuli. In the validation block, their generalization performance was measured with a subset of non-training stimuli without feedback. Day 2 consisted of a refresher stage and an evaluation stage. In the refresher stage, participants completed several training blocks. In the evaluation stage, participants completed the treasure hunt task and the 2AFC task with all stimuli included and no feedback provided. We introduced the stimuli and tasks used in different stages of the experiment as follows.

4.2.1 Stimuli generation and stimuli subsets

We generated new stimuli from the factorial combinations of two sets of symbols. Each stimulus has a left component and a right component. The left component comes from one set of symbols, while

the right component comes from another set of symbols (Figure 4-2A, example of animal symbols in one map). Participants learn to predict the location in space based on components of a stimulus. The hidden rule was that each component of the cue (left or right) indicated the treasure location on the x- or y-axis in Cartesian coordinates. The mapping from left/right to x-/y-axis was fixed for each participant and was arbitrary and randomised across participants.

In the first pilot, we used two symbol sets with five in each, yielding 25 stimuli occupying 25 locations in total. From the second pilot on, we used four symbol sets (X1, X2 and Y1, Y2) to generate four maps (training maps X1Y1, X2Y2 and cross maps X1Y2, X2Y1) with 25 stimuli in each map. To avoid poor performance arising from the inability to remember or distinguish components, we used familiar symbols such as Greek letters (in the one-map pilot) or animals (in all multi-map pilots).

Consistent with the fMRI study, we picked the axis-aligned blocked training curriculum (Chapter 2, Figure 2-1C). In this curriculum, the training stimuli were selected from locations aligned to the cardinal axes and constituted a "+" in the centre of the map, i.e., stimuli in the middle row and column. A main difference was that the stimulus at the centre location was no longer included in the training. During pre-training, training stimuli were presented in a blocked fashion in each training block. Participants first saw stimuli on one axis, and then stimuli on the other axis. The order in which the axes were presented alternated across training blocks. To avoid over-familiarisation, we split the remaining non-training stimuli into two subsets: *validation* stimuli and *test* stimuli (Figure 4-2B). The split was arbitrary and randomised across participants. In the pre-training stage (Day 1), participants' generalization performance was measured with the validation stimuli.

In summary, there existed three stimulus groups (Figure 4-2B) presented at different stages of the experiments (Figure 4-2E):

Training stimuli: stimuli that were used to train participants on the rules. Feedback about the correct treasure locations of training stimuli was provided during the pre-training and refresher stage of the treasure hunt task.

Validation stimuli: a subset of the non-training stimuli that were used during the pretraining stage to measure generalization performance. No feedback was provided about the correct locations of validation stimuli.

Test stimuli: the remaining non-training stimuli that were not seen until the evaluation stage. No feedback was provided about the correct locations of the test stimuli.

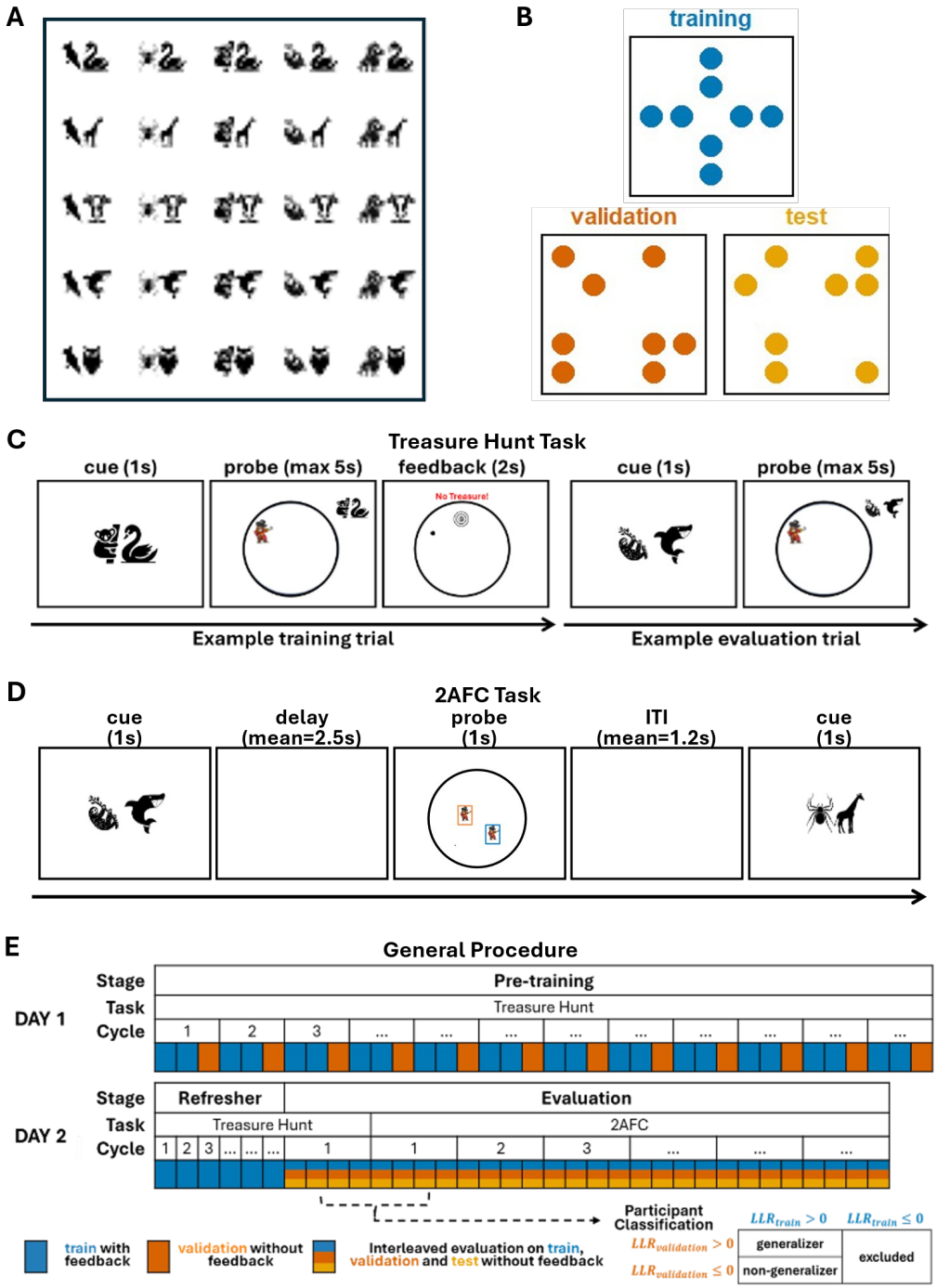


Figure 4-16. Stimuli Set and Task Pool. (A) Example stimuli set and ground-truth map (B). Stimuli groups: example of training, validation and test stimuli. (C) Treasure hunt task trial timeline. Left: timeline of training trial with feedback. Right: timeline of evaluation trial without feedback. (D) 2-Alternative-Forced-Choice (2AFC) task trial timeline. (E) General procedure of the two-day experiment.

4.2.2 Task Pool

4.2.2.1 Tresure hunt task

The treasure hunt task was the same as in the fMRI study in the previous chapter, except for a change in response mode. Instead of using the keyboard, participants clicked on different places on the arena to place the pirate. In each trial (Figure 4-2C), participants first saw a cue stimulus appear in the centre of the screen for 1 second. Then, a circular arena appeared, and participants had to indicate the treasure location with mouse clicks. A pirate will appear at the clicked location and participants can adjust the pirate position with clicks until they decided to submit their response by pressing the space bar. Participants were allowed a maximum of 5 seconds to make a response. Feedback was only provided in the training trials during pre-training and refresher.

We included attention check trials in the treasure hunt task for participant exclusion. Each cycle had one attention check trial. In the attention check trial, participants saw a cue stimulus appear in the centre of the screen for 1 second, as in the task trials. Then, instead of moving the pirate to hunt treasure, they were asked to choose which of six possible stimuli had just been presented. We calculated the mean accuracy across all attention check trials and excluded participants whose accuracy fell below 50%.

4.2.2.2 Two-alternative forced choice (2AFC) task

The 2AFC task was designed to replace the treasure hunt task in the scanner. In this task, participants had to choose the correct treasure location associated with a stimulus within a limited amount of time (Figure 4-2D). In each trial, participants first saw a cue stimulus appear in the centre of the screen for 1 second. After a jittered delay (duration sampled from a Poisson distribution, mean = 2.5s, range = [1s,6s]), participants saw a probe screen displaying two options,

represented by two pirates positioned in different locations within the arena. Participants were allowed a maximum of 1 second to choose the option they believed to be the treasure location associated with the cue stimulus. No feedback was provided for their response. Participants were instructed to think about the treasure location once they saw the cue stimulus, which would help them respond faster once they saw the options on the probe screen. From an fMRI task design perspective, this would help elicit a stronger inference-related signal during the cue screen. We also introduced the jittered delay and jittered intertrial interval (ITI, duration sampled from a Poisson distribution, mean = 1.2s , range = [1s,6s]) to mimic the timeline of our potential fMRI task design.

On the probe screen, the two options were highlighted with blue and orange boxes, respectively. Participants were instructed to press ‘F’ if the blue pirate occupied the treasure location and press ‘J’ if the orange pirate occupied the treasure location. Given that there were five x-axis locations and five y-axis locations, there were 25 possible locations on the arena (regardless of the number of maps). Blue and orange pirates could appear in any of the 25 possible locations. We referred to the correct option as the target and the incorrect option as the distractor hereafter. In half of the trials, the blue option was the target, and in the other half, the orange option was the target. We calculated the distance between the target and the distractor as the measure for trial difficulty. The choice of target-distractor combinations will be explained separately for each pilot experiment below.

Before the 2AFC task, participants completed a minimum of 20 practice trials to familiarise themselves with the key presses. The practice trials had the same timeline as the 2AFC task. The only difference was that on the cue screen, participants saw a pirate

appear at a location on the arena instead of a stimulus. We computed the practice accuracy with a sliding window of 10 trials, and participants had to reach 80% accuracy before they could proceed to the 2AFC task.

4.2.2.3 Symbol arena task

After the evaluation stage and at the end of the experiment, we asked participants to complete a task to see if we could probe their internal representation of rules. All symbols that made up the stimuli were presented on the screen (randomly placed on a rectangular canvas). Participants were asked to drag and drop the symbols to arrange them in a way that represented their relationships to one another.

4.2.3 Performance Metrics

4.2.3.1 Performance metrics of the treasure hunt task

The same performance metrics as the previous fMRI chapter were used for the treasure hunt task: distance to goal and log-likelihood ratio (LLR). Distance to goal was defined as the Euclidean distance between a participant's response location and the ground-truth location in a trial, i.e. a metric for response error. LLR compared whether a participant's response in a given treasure hunt task trial was better described by a ground-truth model or a null model. The ground-truth model assumed that the response would follow a bivariate binormal distribution centred around the true treasure location. In contrast, the null model assigned equal response likelihood across the arena. Details on the calculation of these metrics can be found in Chapter 3, Section **2.3.5 Participant classification and exclusion**.

4.2.3.2 Performance metrics for the 2AFC task

Two performance metrics were used for the 2AFC task: choice accuracy and reaction time (RT). Trials where participants failed to

make a choice within the time limit (1s) were considered missed trials and were discarded from the analysis. If the choice was the target option, the trial would count as correct. While we instructed participants to think about the treasure location once they saw the cue stimulus, the options were not available until the probe screen. Therefore, we defined reaction time as the time interval between the onset of the probe screen and the moment the choice was made.

4.3. Pilot M1V1: one map task with the symbol combination stimuli

Before we ran the multi-map experiments, we first validated that participants were still able to work out the compositional generalization rules with the new stimuli. Therefore, we conducted a small pilot experiment with only one map to see if participants could generalize as in the previous animal-shape experiments.

4.3.1 Methods

4.3.1.1 Participants

9 participants recruited on Prolific completed the experiment.

4.3.1.2 Procedure

The experiment consisted of two days (Figure 4-3A). Day 1 was the pre-training stage, where participants completed 12 train-validation cycles. Each cycle had two training blocks followed by a validation block. Day 2 started with a refresher stage where participants completed six training blocks with feedback. Then, participants completed the evaluation stage. First, they did one cycle of the treasure hunt task without feedback. Then, they completed six cycles of the 2AFC task. During the evaluation stage, all stimuli were

presented four times in this cycle. This was to keep the cycle length consistent with the multi-map design.

4.3.1.3 Design

We selected ten Greek letters and split them into two sets with five in each. The split was random for each participant. We generated stimuli for this one-map pilot from the factorial combination of these two sets, excluding the central one at (0,0). This yielded 24 stimuli (Figure 4-3B). As described in the general methods section, stimuli in the central row and column were the training stimuli, and the remaining 16 stimuli were then split into eight validation stimuli and eight test stimuli (Figure 4-2B). This split was also random for each participant.

In the two-AFC task, we sampled all possible combinations of targets and distractors. In total, there were 25 locations on the map (the central one was taken into account as well). As we only had one map for this pilot, the number of distractors for each possible target stimulus was 24*1=24. We had six cycles of the 2AFC task with four repeats per stimulus in each cycle, yielding 24 repeats per stimulus in total. This eventually allowed us to sample each possible target-distractor combination once.

The trial timelines for the treasure hunt task and the 2AFC task were described in the previous general methods section (Figure 4-2 C&D).

4.3.1.4 Analysis

Participant classification according to treasure hunt task performance

Similar to the previous chapter, we defined participants who could correctly solve the training stimuli as learners and those who failed as non-learners. We further defined participants who could correctly

solve the non-training stimuli as generalizers and those who failed as non-generalizers.

Importantly, we classify participants based on their performance in the validation stimuli trials in the treasure hunt task during the evaluation stage. For each participant's response during the evaluation stage of the treasure hunt task, we computed the average LLR of training stimuli (LLR_{train}), validation stimuli ($LLR_{validation}$) and test stimuli (LLR_{test}) separately. We further excluded participants with $LLR_{train}<0$ (non-learners). The remaining participants with $LLR_{train}>0$ (learners) were classified as generalizers ($LLR_{test}<0$) or non-generalizers ($LLR_{test}<0$).

Analysis

Given the limited number of participants in this pilot, we did not perform statistical tests at the group level. We plotted individual performance alongside group average performance to describe the trend of the data. For the treasure hunt task, we additionally plotted the averaged performance from the previous fMRI chapter for reference.

For the 2AFC task, we ran regression models separately for the choice accuracy and RT for each participant to examine the impact of stimuli group (training, validation and test), trial difficulty (diff.), as well as trial progression (trial). Test was coded as the reference level for stimuli group and yielded two dummy-coded independent variables that contrasted training and validation stimuli against test stimuli: Train>Test, Val.>Test. As mentioned in the general method section, trial difficulty was defined as the city block distance between the target and the distractor, with increasing distance indicating easier trials. Trial was defined as the trial number across the 2AFC task. Difficulty and trial were continuous variables, so they were

standardised before running the regression. All the interaction terms were included in the regression models. We fitted: 1) a logistic regression model to the choice accuracy of trials in which participants responded within the time limit, and 2) a linear regression model to the RT of correct trials.

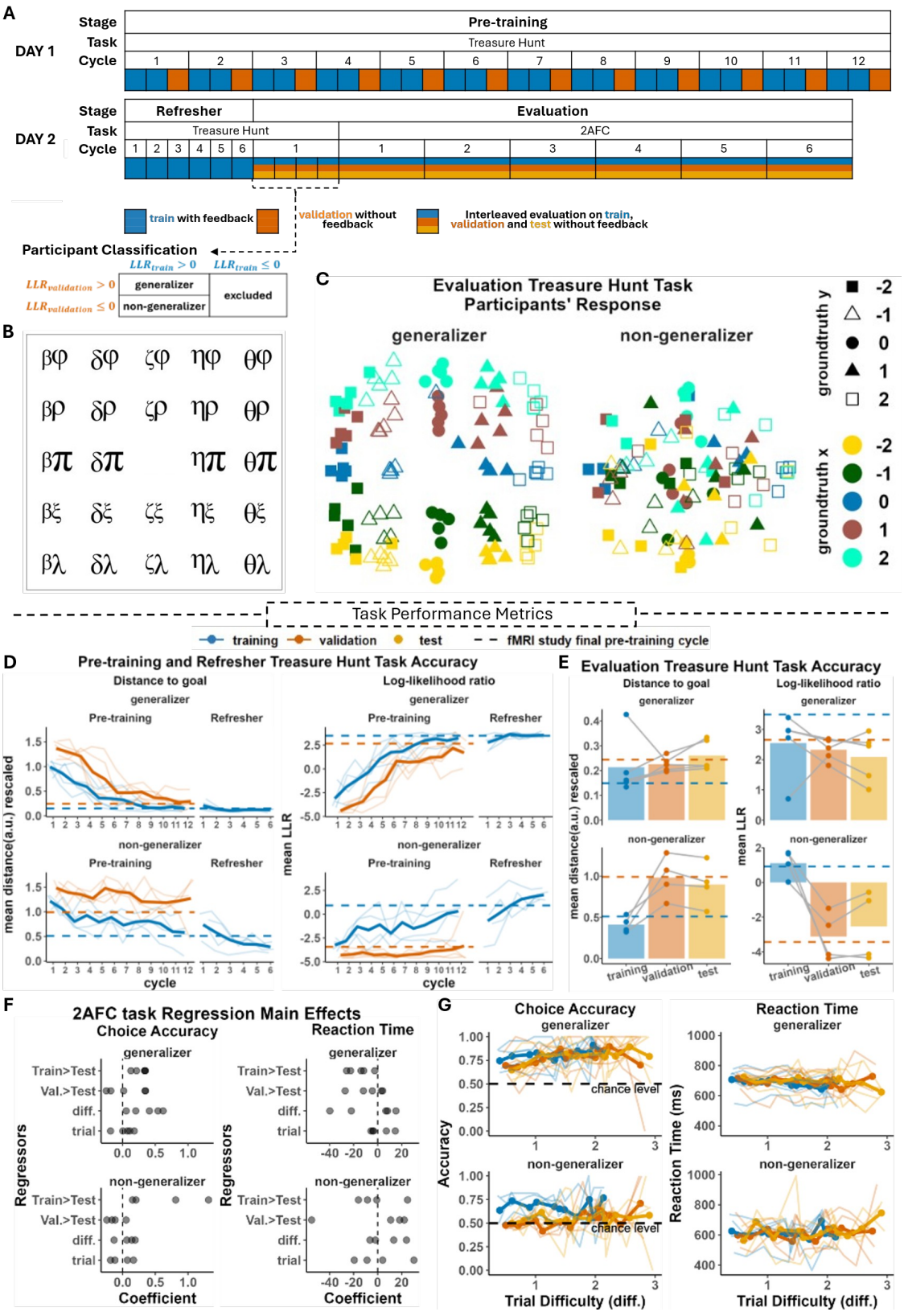


Figure 4-17. Procedure and performance in the one-map pilot (M1V1). (A) Procedure of the one-map pilot (M1V1). (B) Example stimuli and ground-truth map (C) Participants' response map during the evaluation stage on Day 2. (D-E) Participants' performance in the treasure hunt task

was measured by distance to goal and log-likelihood ratio (LLR) during pretraining and refresher (D), as well as (E) during evaluation. Solid lines indicated the performance of individual participants (translucent lines) and average performance (opaque lines). Dashed lines indicated the average performance from the previous fMRI chapter. (F-G) Participants' performance in the 2AFC task: (F) Main effects from the regression on choice accuracy (left) and reaction time (right). Dots indicated the regression coefficients from individual participants. (G) Main effect of trial difficulty (distance between target and distractor) on choice accuracy (left) and reaction time (right). Dashed lines indicated the chance level. Continuous solid lines indicated the performance of individual participants (translucent lines) and average performance (opaque lines).

4.3.2 Results and Discussions

We classified participants based on whether they successfully inferred the treasure locations in the validation stimuli during the evaluation phase of the treasure hunt task (Figure 4-3A). Those who succeeded were classified as generalizers, while those who did not were classified as non-generalizers.

We observed that towards the end of the pre-training stage, generalizers' performance approached the average performance in the fMRI study (Figure 4-3D). This indicated that participants were still able to learn the compositional generalization rule with the new stimulus set generated from symbol combinations.

More importantly, classification based on validation stimuli trials in the treasure hunt task seemed robust. Within the treasure hunt task, participants' performance in the test stimuli trials did not seem to differ from the validation stimuli trials (Figure 4-3E). This ensured that generalizers' performance in the test stimuli (all generalizers' $LLR_{test}>0$) was better than the non-generalizers (all non-generalizers' $LLR_{test}<0$). Relatedly, generalizers' response map largely resembled

the ground-truth map of the stimuli's treasure locations (Figure 4-3C). The classification also seemed robust in the 2AFC task. The mean choice accuracy for training, validation and test stimuli was well above chance for generalizers (Figure 4-3G, top-left panel). In comparison, the mean choice accuracy for validation and test stimuli was all at chance for non-generalizers (Figure 4-3G bottom left panel).

Finally, performance in the 2AFC task was sensitive to the trial difficulty. We defined trial difficulty as the distance between the target and distractor option on the probe screen. Regression analysis on choice accuracy revealed a consistently positive coefficient for trial difficulty for all generalizers (Figure 4-3F), indicating that greater distance (reduced trial difficulty) was associated with increased choice accuracy.

4.4. Pilot M2V1: multi-map task, blocked evaluation

We've established in pilot M1V1 that participants understand and learn the task with the symbol combination stimuli to a similar extent as the fMRI study using the coloured animal stimuli. Therefore, in this pilot (M2V1), we want to see if they can learn two maps and generalize across maps.

4.4.1 Methods

4.4.1.1 Participants

13 participants recruited from Prolific completed the study. We excluded 1 participant who failed the attention check in more than 50% of the trials.

4.4.1.2 Design and procedure

In this chapter, we reported two pilot experiments of the multi-map tasks. As there were not enough visually distinctive Greek letters

available, we moved to use animals to generate combinations in the multi-map tasks (Figure 4-4A). The introduction of multiple training maps (Figure 4-4A, main diagonal) allowed us to test cross-map generalization performance (Figure 4-4A, anti-diagonal). Two cross maps were generated. Recombining the x-axis symbols from training map 1 (T1) with the y-axis symbols of training map 2 (T1) yielded cross map 1 (C1). Similarly, reversing the combination yielded cross map 2 (C2). In the training maps, the partitioning of training, validation and test stimuli subset was the same as in the one-map task with the additional constraints that the validation set of T1 and T2 were complementary to each other (Figure 4-4B) and that the central stimulus in each map would always be in the test subset.

The procedure was also similar to the one-map task except that all the stages were equally split between maps (Figure 4-4C). In the pre-training stage on day 1, the first eight cycles were used to train participants on training map 1 (T1), and the remaining eight cycles were training map 2 (T2). On Day 2, participants first received an equal amount of refresher training on T1 and T2 in a blocked fashion. They completed four cycles of T1 refresher and then four cycles of T2 refresher. In the evaluation stage, participants were evaluated on all the stimuli from both the training maps (T1, T2) and cross maps (C1, C2). We gradually built up the complexity of the evaluation stage across pilot experiments.

In the current first pilot of the multi-map task (pilot M2V1), maps were evaluated in a blocked fashion where participants completed all trials from one map before proceeding to another map (Figure 4-4C). We started with a blocked design for evaluation, as it would be the least confusing for participants. Participants first completed one cycle of the treasure hunt task, where each map was evaluated once, and the training map preceded the cross maps. The idea was that participants

could 1) practice on the within-map generalization before doing the cross-map generalization, and 2) practice on the cross-map generalization in the more familiar treasure hunt task before doing the 2AFC task. In the 2AFC task, there were six cycles, and each map was evaluated once in every cycle. The order of map blocks within each cycle was randomised regardless of training or cross maps.

Finally, as there were 100 stimuli in total, each stimulus could only be repeated six times in the 2AFC task. It was not possible to sample all possible target-distractor combinations (which required 24 repeats per stimulus). Pilot M1V1 showed that the 2AFC task performance was sensitive to trial difficulty, i.e., the distance between the target and the distractor. Therefore, we sampled the distractor according to different levels of difficulty. More specifically, we split all possible distractors (24 in total) into three categories according to their city block distance to the target stimulus: one-step, two-step, and ≥3-step. We then sampled two trials from each category for each target stimulus.

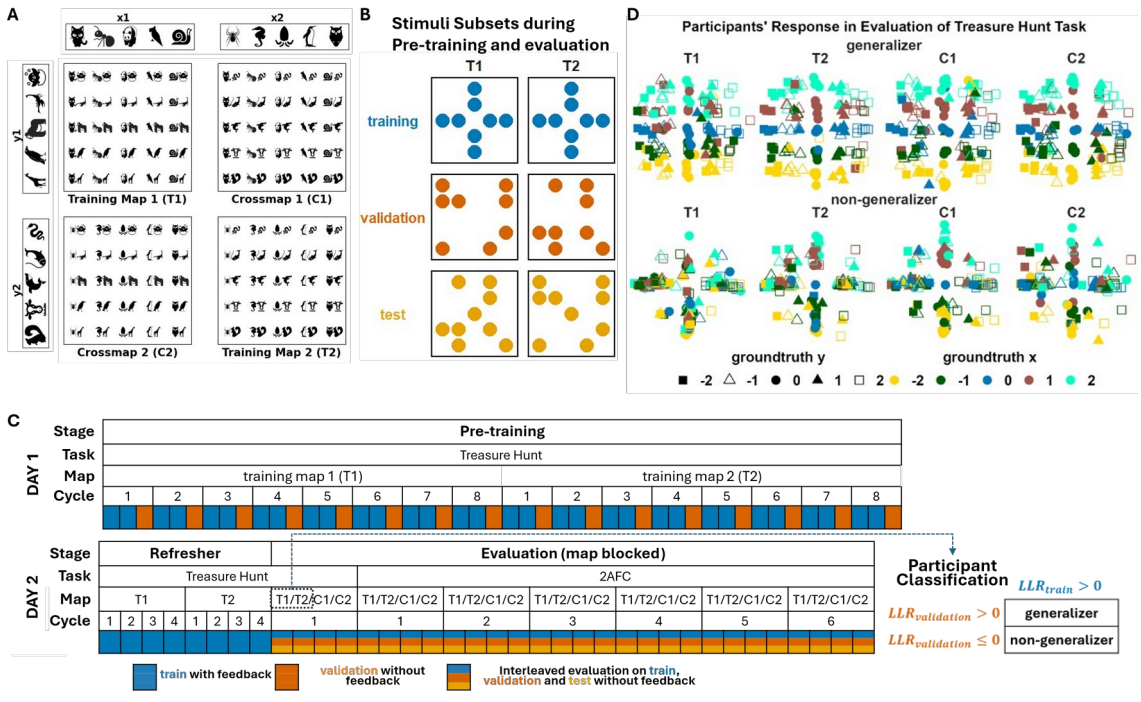


Figure 4-18. Procedure and response map in the first multi-map pilot (pilot M2V1). (A) Stimuli and the ground-truth locations of different training maps (main diagonal) and different cross maps (anti-diagonal). (B) Stimuli subset on the training maps. (C) Procedure of M2V1. (D) Participants' response map in the evaluation of the treasure hunt task on day 2.

4.4.1.3 Participant classification

Similar to pilot M1V1, participants were classified according to their performance in the evaluation stage of the treasure hunt task (Figure 4-4C). More specifically, we computed the average LLR of the training stimuli trials across both training maps (LLR_{train}), and the average LLR of the validation stimuli trials across both training maps ($LLR_{validation}$) for each participant. Participants with $LLR_{train}<0$ were excluded and the remaining participants were classified as generalizers ($LLR_{test}<0$) or non-generalizers ($LLR_{test}<0$).

4.4.1.4 Analysis

Given the limited number of participants in this pilot, we did not perform statistical tests at the group level. We plotted individual performance alongside group average performance to describe the trend of the data. For the treasure hunt task, we additionally plotted the averaged performance from the previous fMRI chapter for reference.

4.4.2 Results and Discussions

Participants were classified into generalizers and non-generalizers based on their performance in the validation stimuli trials in the treasure hunt task during evaluation (Figure 4-4C).

While participants only received half the amount of pretraining and refresher on each training map in the current pilot compared to the

one-map pilot M1V1, their performance still approached the average performance in the fMRI study (Figure S4-6). This indicated that participants were still able to learn the compositional generalization rule in the multi-map task with stimuli generated from animal symbol combinations. Participants on average missed 5.9% of trials in the 2AFC task in the current pilot, which was comparable to the one-map pilot (Figure S4-2, Table S4-1). Additionally, we also observed a transfer learning effect, whereby participants learned T2 faster than T1 (Supplementary results 8.5).

Similar to pilot M1V1, participants' performance on the validation stimuli and test stimuli was comparable (Supplementary results 8.3). Generalizers' response maps largely resembled the ground-truth map of the stimuli's treasure locations (Figure 4-4D top panels). Interestingly, while the non-generalizers' response in non-training stimuli trials failed to match the ground-truth locations in most cases, their response in trials with stimuli on the central row/column in the cross maps seemed consistent with the ground-truth (Figure 4-4D bottom panels, see Supplementary results 8.3 for further analysis on error). This observation led us to re-classify the stimuli as centre vs non-centre stimuli across all maps (Figure 4-5A). Indeed, when plotting the response error separately for centre and non-centre stimuli, we can see that numerically, non-generalizers responded more accurately in trials with centre stimuli than with non-centre stimuli (Figure 4-5B). This was true for the cross maps, even though they never received feedback about the correct locations of stimuli from the cross maps. In the 2AFC task, this differential response pattern between centre vs. non-centre stimuli in the current pilot resembled the pattern of training vs. non-training stimuli in the one-map pilot (M1V1). All generalizers were above chance on both the

centre as well as the non-centre stimuli, whereas most non-generalizers were at chance for the non-centre stimuli (Figure 4-5C).



Figure 4-19. Behavioural performance in the first multi-map pilot (pilot M2V1) during the evaluation stage. (A) schematic illustration showing the centre versus non-centre stimuli dichotomy. (B) Treasure hunt task performance measured by distance to correct location (top) and log-likelihood ratio (bottom). Dots indicated individual participants. Bars indicated the group average. The average performance from the fMRI study in the previous chapter was shown in dashed lines. (C) 2AFC task accuracy (top) and reaction time (bottom). Black dashed lines indicated the chance level. Continuous solid lines indicated individual participants (translucent lines) and group average (opaque lines). We additionally highlighted two non-generalizers who were well above chance with dotted translucent lines.

It is worth highlighting that the current classification seemed to misclassify two of the participants as non-generalizers. We can see from the individual lines in Figure 4-5C that two of the non-generalizers showed above-chance accuracy for both centre and non-centre stimuli (Figure 4-5C top, dotted translucent lines). One of them showed inconsistent behaviour in the treasure hunt task and the 2AFC task. The other non-generalizer, however, showed consistent behaviour across tasks (see Supplementary Results 8.5 for details). This raised an interesting question: if the behaviour was consistent,

why did low average performance in the treasure hunt task lead to high average performance in the 2AFC task?

We examined the response map of this non-generalizer. As it turned out, they failed to remember some of the symbol-location associations, which led to a low mean LLR for test stimuli. This pointed to a potential limitation of the current classification method: its inability to distinguish between mere memory of the rules and the broader capacity for compositional generalization. To be classified as a generalizer under the current method, one had to remember the rules accurately and compose them to make the inference. However, it is possible to have inaccurate memory but still generalize compositionally. We illustrated this using the data from this non-generalizer in Figure 4-6. Compositional generalization indicated that the response in a test trial was generated by combining the response in the relevant training stimuli trials. We took the participant's response during the refresher stage (i.e., training stimuli response, Figure 4-6B), which reflected memory of the rules, and composed them into the putative response to non-centre stimuli (Figure 4-6C). Then, we regressed the participant's observed response during the evaluation stage on the putative response. If a participant generalized compositionally, then the putative response on both the x and y axes should predict the observed response well. After repeating this analysis for each participant, we can see that this non-generalizer was more similar to the generalizers and did show the signature of compositional response (Figure 4-6D), even though the memory of the rules was inaccurate (compare panels A and B in Figure 4-6). Remembering rules accurately was less of an issue in the previous fMRI experiment, as there were only 10 feature-location associations to learn. The inherently factorised colour-vs-shape dimension also helped participants separate x-axis associations and y-axis

associations. In the multi-map task, however, the number of feature-location associations was doubled. Additionally, participants had to memorise which symbols corresponded to the x-axis locations, and which corresponded to the y-axis locations. This increased memory demand made the original classification method no longer suitable. In the next pilot, we explored a new classification method that aimed to address this concern.

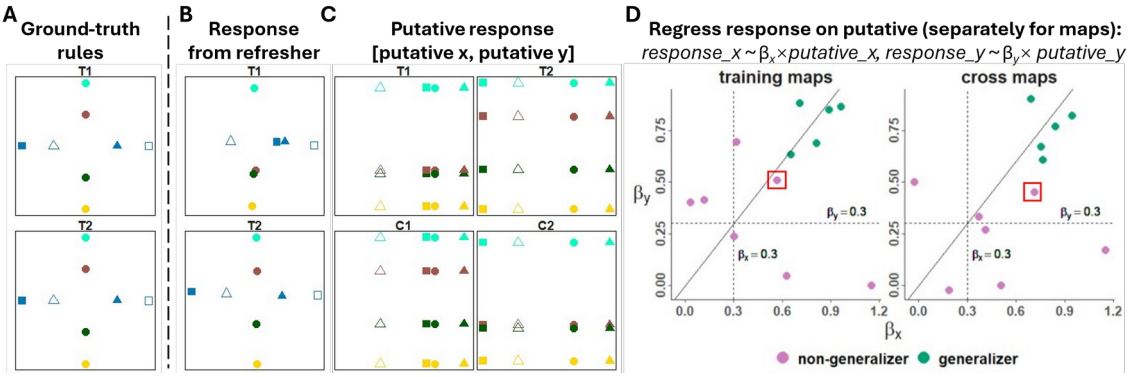


Figure 4-20 . Illustration of compositional generalization with inaccurate rule memory. (A) ground truth rules. (B) Participants' response from the refresher stage reflected their memory of the rules. (C)The putative response to all non-centre stimuli was generated by combining the refresher stage response. (D) If a participant responded with a compositional strategy, their observed response should correlate with the putative response. The example non-generalizer classified by the current method (highlighted by a red rectangle) showed the signature of compositional generalization.

4.5 Pilot M2V2 and pilot M2V3: multi-map task, randomised evaluation, new classification method

We established in pilot M2V1 that participants understand and learn the task and generalize across maps. We also recognised potential

pitfalls of the previous classification method. In this pilot, we wanted to remove the blocked design during the evaluation stage and randomise the stimulus presentation in a cycle across maps, as it would be optimal for fMRI study design. Additionally, we developed an improved classification method.

More importantly, we modified the participant classification procedure to address the limitations highlighted in pilot M2V1.

4.5.1 Methods

4.5.1.1 Participants

39 participants recruited on Prolific completed pilot M2V2, 4 were excluded from the following analysis because they failed the attention check in more than 50% of trials. In pilot M2V3, 34 participants recruited on Prolific completed the experiments, and 5 were excluded according to the same criteria. We combined the data from both pilots in the following result section (N=64).

4.5.1.2 Design and procedure

The procedure was almost identical to pilot M2V1 (see section 4.1.2 and Figure S4-1), except that:

During the refresher, participants completed the refresher of T1 and T2 in an alternating fashion. In other words, participants completed one block of training on T1 (8 trials) and then one block of training on T2 (8 trials), and this T1-T2 refresher cycle was repeated four times.

In each cycle of the evaluation stage, presentation of stimuli was randomised across maps. In other words, all 100 stimuli from the four maps (two training maps and two cross maps) were presented in random order in an evaluation cycle.

Additionally, in pilot M2V3, we switched the order of tasks during the evaluation stage, i.e., participants performed the 2AFC task before the treasure hunt task.

4.5.1.3 New classification method

Framework

The new classification method leveraged a model comparison approach to determine whether a participant's responses in the test stimulus trials were best explained by a model that solved the task compositionally (referred to as ***the compositional model*** hereafter). Under the compositional model, a participant's response to a given test stimulus (e.g., *snail-chicken* in Figure 4-4A, training map 1) should be derived from their responses to two relevant training stimuli (e.g., *snail-chimpanzee* and *panda-chicken* in the same map). Specifically, we took the participant's response on the x-axis from the *snail-chimpanzee* trial and their response on the y-axis from the *panda-chicken* trial. These values were referred to as ***the putative response*** for the x- and y-axes of the *snail-chicken* test trial. The compositional model assumed that the observed responses on both axes were drawn from Gaussian distributions centred around the putative responses. This rationale closely mirrors the illustration in Figure 4-6. However, instead of using regression to assess the correlation between the putative and observed responses, we employed model comparison to determine whether the compositional model was the winning model. In the next part, we will describe how this compositional model and other control response models were set up.

The central idea behind the models was to characterise participants' responses as being drawn from different probability distributions: $p(x, y|model)$ and to find out the distribution that best described the

data. First, let us consider ***the response strategy*** for each axis: $p(x|x_{strategy})$ and $p(y|y_{strategy})$. Participants made their responses within a circular arena of radius r. The observed responses on the x-axis and the y-axis were denoted as x and y, and the corresponding putative responses as $x_{putative}$ and $y_{putative}$. We considered three different strategies: *putative*, *biased* or *random*. The *putative* strategy indicated that the response likelihood would be highest near the putative locations. The *biased* strategy reflected a tendency to respond in a specific location on the axis, regardless of the stimulus. The *random* strategy assumed that responses were equally likely at any location on the axis (i.e., a uniform distribution). Take the x-axis as an example, the three strategies can be written down as the following mixtures of distributions:

putative:

$$p(x|x_{putative})=(1-\alpha)\times N(x_{putative},\sigma_x)+\alpha\times U(-b,b)$$ *Equation 1*

bias:

$$p(x|x_{bias})=(1-\alpha)\times N(x_{bias},\sigma_x)+\alpha\times U(-b,b)$$ *Equation 2*

random:

$$p(x|x_{random})=U(-b,b)$$

Equation 3

$N(\mu,\sigma)$ was a Gaussian distribution with a mean of μ and a variance σ^2. $U(-b,b)$ was a uniform distribution bounded by $[-b,b]$. b was computed from the arena radius and the response on the y-axis: $b=\sqrt{r^2-y^2}$. α was a 1D lapse rate controlling the randomness of response on the x-axis. Similar equations can be derived for the y-axis response.

Response on the x-axis and the y-axis can have different strategies. Assuming that responses on different axes were drawn independently, a multiplicative rule was used to derive the probability distribution of the observed response vector (x, y) following a certain strategy combination $x_{strategy}$ - $y_{strategy}$:

$$p(x, y | x_{strategy}, y_{strategy}) = p(x | x_{strategy}) \times p(y | y_{strategy}) \qquad \textit{Equation 4}$$

Finally, a model may reflect *a collection of strategy combinations*. In a collection of K strategy combinations ($K \geq 1$), the participant would choose a strategy combination randomly according to some weight, and we used w_k to denote the weight assigned to the kth strategy combination. Taken together, the likelihood of the observed response x, y in trial t under a given model was computed as:

$$L_t = p(x, y | model)$$

$$¿(1-\beta) \times \sum_{k=1}^{K} w_k \times p(x, y | x^k_{strategy}, y^k_{strategy}) + \beta \times \frac{1}{\pi r^2} \qquad \textit{Equation 5}$$

β was a 2D lapse rate introduced to account for randomness in the overall response pattern.

In summary, we set up a framework for building response models from a collection of x-y strategy combinations. Under this framework, a certain x-y strategy combination specified a joint probability distribution from which an observed response vector (x, y) was drawn. A response model reflected a collection of strategy combinations that were combined linearly with some weights. Given this nature of linear combination of distributions, models built under this framework fell within the family of mixture models.

Model space

Now that we've described our framework for building models from a collection of x-y strategy combinations, we move on to describe the

specific models we used for the current pilot study. Note that we reduced the model space (by fixing some parameters and limiting the strategy combinations) based on empirical observations, but future work could consider the entire model space under this framework. From now on, we will denote any strategy combination as $x_{strategy}$-$y_{strategy}$, e.g. *putative-random* meant that the response on the x-axis followed the *putative* strategy and the response on the y-axis followed the *random* strategy.

Several parameters were fixed or constrained:

When a model specified a collection of strategy combinations, each combination was assumed to be equally likely. In other words, if a model specified a collection of K strategy combinations, we set $w_k = \frac{1}{K}$.

All (non-random) strategies were assumed to have a shared 1D lapse rate.

We considered four models (Figure 4-7). The first two models (null and compositional) were similar to the previous classification method (see **2.3.5 Participant classification and exclusion**, Figure 2-2B, null and ground-truth model). Apart from estimating the variances of the distribution as free parameters, the compositional model replaced the ground-truth locations with putative locations.

In the null model, there were no strategy combinations. Participants were responding entirely randomly, i.e. the model assigned equal probability to all locations within the circular arena. This essentially set β as 0 in Equation 5:

$$p(x, y|null\,model) = \frac{1}{\pi r^2}$$

There were no free parameters in this model.

The compositional model

The composition model only had one strategy combination: putative-putative, meaning that the responses on both x- and y-axes should come from distributions centred around the putative locations.

$$p(x,y|compositional\,model)=(1-\beta)\times p(x,y|x_{putative},y_{putative})+\beta\times\frac{1}{\pi r^2}$$

$$¿(1-\beta)\times\left[p(x|x_{putative})\times p(y|y_{putative})\right]+\beta\times\frac{1}{\pi r^2}$$

where

$$p(x|x_{putative})=(1-\alpha)\times N(x_{putative},\sigma_x)+\alpha\times U(-b_x,b_x)$$

$$p(y|y_{putative})=(1-\alpha)\times N(y_{putative},\sigma_y)+\alpha\times U(-b_y,b_y)$$

The bounds of the uniform distribution were calculated based on the arena radius and the response on the other axis: $b_x=\sqrt{r^2-y^2}$ and $b_y=\sqrt{r^2-x^2}$. Therefore, this model had four free parameters, i.e., the lapse rates α and β as well as the variances for the Gaussians σ_x and σ_y.

Across the many experiments we have conducted with the pirate task, we noticed that the response patterns of non-generalizers were not random either. Their responses exhibited partial alignment with the putative locations, such that the response matched the putative value on one axis but deviated on the other. Therefore, we set up two 1D models to capture this observation, the response on one axis followed the putative strategy, but was random or biased on the other.

The 1D-centre model:

Under the 1D centre model, the participants followed the putative strategy on one axis but showed centrality bias on the other. Hence, it

reflected a collection of two strategy combinations: *putative-bias* or *bias-putative*.

$$p(x,y|1Dcentre)$$

$$¿(1-\beta)\times\left[\frac{1}{2}\times p(x,y|x_{putative},y_{bias})+\frac{1}{2}\times p(x,y|x_{bias},y_{putative})\right]+\beta\times\frac{1}{\pi r^2}$$

$$¿(1-\beta)\times\frac{1}{2}\times\left[p(x|x_{putative})\times p(y|y_{bias})+p(x|x_{bias})\times p(y|y_{putative})\right]+\beta\times\frac{1}{\pi r^2}$$

where $p(x|x_{putative})$ *and* $p(y|y_{putative})$ *were calculated in the same way as in the compositional model, and*

$$p(x|x_{bias})=(1-\alpha)\times N(x_{bias},\sigma_{biasx})+\alpha\times U(-b_x,b_x)$$

$$p(y|y_{bias})=(1-\alpha)\times N(y_{bias},\sigma_{biasy})+\alpha\times U(-b_y,b_y)$$

In theory, x_{bias} and y_{bias} could be free parameters. Here, we only considered the centrality bias by setting $x_{bias}=y_{bias}=0$ based on empirical observations in the previous experiments that participants who failed to generalize tend to put things along the central axes. Finally, the bounds of the uniform distribution are calculated in the same way as the compositional model. Therefore, this model had six free parameters, i.e., the variances for the Gaussians σ_x and σ_y in the putative strategies, the variances for the Gaussians σ_{biasx} and σ_{biasy} in the bias strategies, as well as the lapse rates α and β.

The 1D-random model

Under the 1D-random model, the participant followed the putative strategy on one axis but was random on the other. Hence, it reflected a collection of two strategy combinations: *putative-random* or *random-putative*.

$$p(x,y|1Drandom)$$

$$¿(1-\beta)\times\left[\frac{1}{2}\times p(x,y|x_{putative},y_{random})+\frac{1}{2}\times p(x,y|x_{random},y_{putative})\right]+\beta\times\frac{1}{\pi r^2}$$

$$¿(1-\beta)\times\frac{1}{2}\times\left[p\left(x\middle|x_{putative}\right)\times p\left(y\middle|y_{random}\right)+p\left(x\middle|x_{random}\right)\times p\left(y\middle|y_{putative}\right)\right]+\beta\times\frac{1}{\pi r^2}$$

where $p\left(x\middle|x_{putative}\right)$ *and* $p\left(y\middle|y_{putative}\right)$ *were calculated in the same way as in the compositional model, and*

$$p\left(x\middle|x_{random}\right)=U\left(-b_x,b_x\right)$$

$$p\left(y\middle|y_{random}\right)=U\left(-b_y,b_y\right)$$

Again, the bounds of the uniform distribution were calculated from the arena radius and the response on the other axis. Therefore, the model had four free parameters, i.e., the variances for the Gaussians σ_x and σ_y in the putative strategies and the lapse rates α and β.

We simulated responses under the four specified models (Figure 4-7). 100 simulations were performed using the following parameters:

Compositional model: $\sigma_x=\sigma_y=0.05,\alpha=\beta=0$

1D-center model: $\sigma_x=\sigma_y=\sigma_{biasx}=\sigma_{biasy}=0.05,\alpha=\beta=0$

1D-random model: $\sigma_x=\sigma_y=0.05,\alpha=\beta=0$

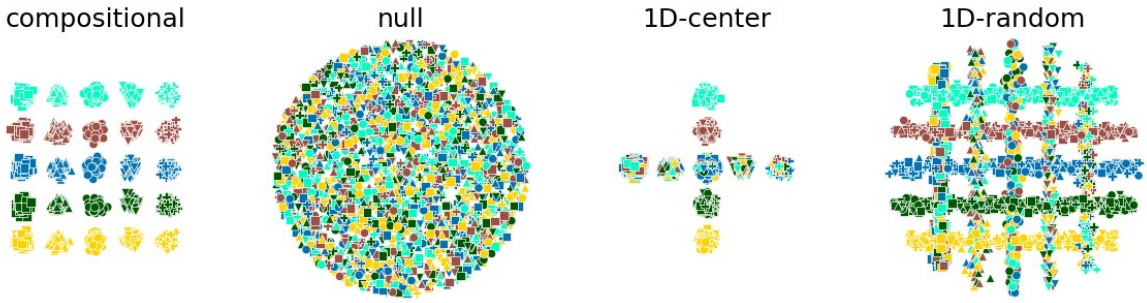


Figure 4-21. Model Space. We simulated responses under the four specified models. 100 simulations were generated for each model. Each dot represented the simulated response to one stimulus. Shapes indicated the ground-truth of x locations, and colour indicated the ground-truth of y locations.

Model fitting, model comparison and cross-validation

For each participant, we fitted all four models using their data from trials containing the validation stimuli in the training maps (16 stimuli in total, referred to as '*fitting data*' hereafter). Participants were then classified based on their winning model, i.e., the model with the smallest Bayesian Information Criterion (BIC). We chose BIC over the raw NLL to penalise for over-parameterisation:

$$BIC = NLL + \log(N) * m$$

where N was the total number of trials and m was the total number of free parameters in the model.

Parameters were bounded. The variance parameters for the Gaussians were bounded between 0 and the error tolerance of the task. During training, feedback was given as three concentric circles (corresponding to three levels of bonus). Responses falling within these circles would receive the corresponding level of bonus. The radius of the outermost circle was defined as the error tolerance. The lapse rate parameters were bounded between 0 and 0.1.

The 'minimise' function from the SciPy 'optimize' module was used to find the parameters that minimised the negative loglikelihood (NLL) loss. Given that our parameters were bounded, the Broyden-Fletcher-Goldfarb-Shanno with bounds (L-BFGS-B) algorithm was chosen. Warm starts were used to determine some of the starting values. For the variance parameters, we computed the starting values by computing the standard variance from the observed data. For example, starting values of the variance parameters for the x-axis were computed as $\sigma 0_x = \sqrt{\frac{1}{N} \times \sum (x - x_{putative})^2}$ and $\sigma 0_{biasx} = \sqrt{\frac{1}{N} \times \sum (x - 0)^2}$. Starting values of the lapse rates were randomly drawn from a uniform distribution bounded between 0 and 0.1.

To make sure that we did not overfit to data from trials with the validation stimuli in the training maps, we performed cross-validation (CV). There were four subsets of data not used for model fitting:

Centre stimuli in the training maps (18 stimuli in total)

Centre stimuli in the cross maps (18 stimuli in total)

Non-centre test stimuli in the training maps (16 stimuli in total)

Non-centre test stimuli in the cross maps (32 stimuli in total)

Given the results in pilot M2V1, we were interested in the non-centre stimuli for measuring generalization performance. Therefore, we only used the two non-centre stimuli subsets as CV data.

Model fit on the held-out was evaluated with two metrics: the NLL and the R-squared. We obtained the best-fitting model and the fitted parameters of this model for each participant. Then, we generated the probability of the observed response under the winning model to compute NLL. We also simulated responses from the winning model. This was straightforward as the winning model specified a response probability distribution. Simulating responses was essentially drawing samples from this probability distribution. The R-squared between the simulated response and the observed response can be computed. For each participant, we performed 500 simulations and computed the average R-squared across the simulations.

4.5.2 Results

4.5.2.1 Consistency of the winning models

We fitted the four specified models (Figure 4-7) to the data from trials containing validation stimuli in the training maps. Model comparison was based on the Bayesian Information Criterion (BIC), with the model having the lowest BIC considered the winning model.

Participants whose winning model was the compositional model were classified as compositional; all others were classified as non-compositional.

We first examined whether the model comparison procedure yielded consistent results between the fitting and cross-validation (CV) data. To assess this, we plotted the correlation between model fits in the fitting and CV data (Figure 4-8A). Most participants, particularly those classified as compositional, fell near the diagonal line, indicating that the winning model performed similarly in both datasets. Next, we calculated the BIC for each model on the CV data without refitting to the data, and selected the winning model based on these BIC scores. The results showed that, for the majority (22/30) of compositional participants, the winning model in the CV data matched the one from the fitting data (Figure 4-8B). Together, these results suggested that our new classification procedure yielded consistent results. Note that the model fitting consistency was also worse in non-compositional participants. This could be due to two reasons. It could be that the performance of non-compositional participants was noisier and less consistent than the compositional participants. It could also be that our current model space is limited and has failed to cover the behavioural strategies of some of the non-compositional participants. We explored the second possibility in the next section.

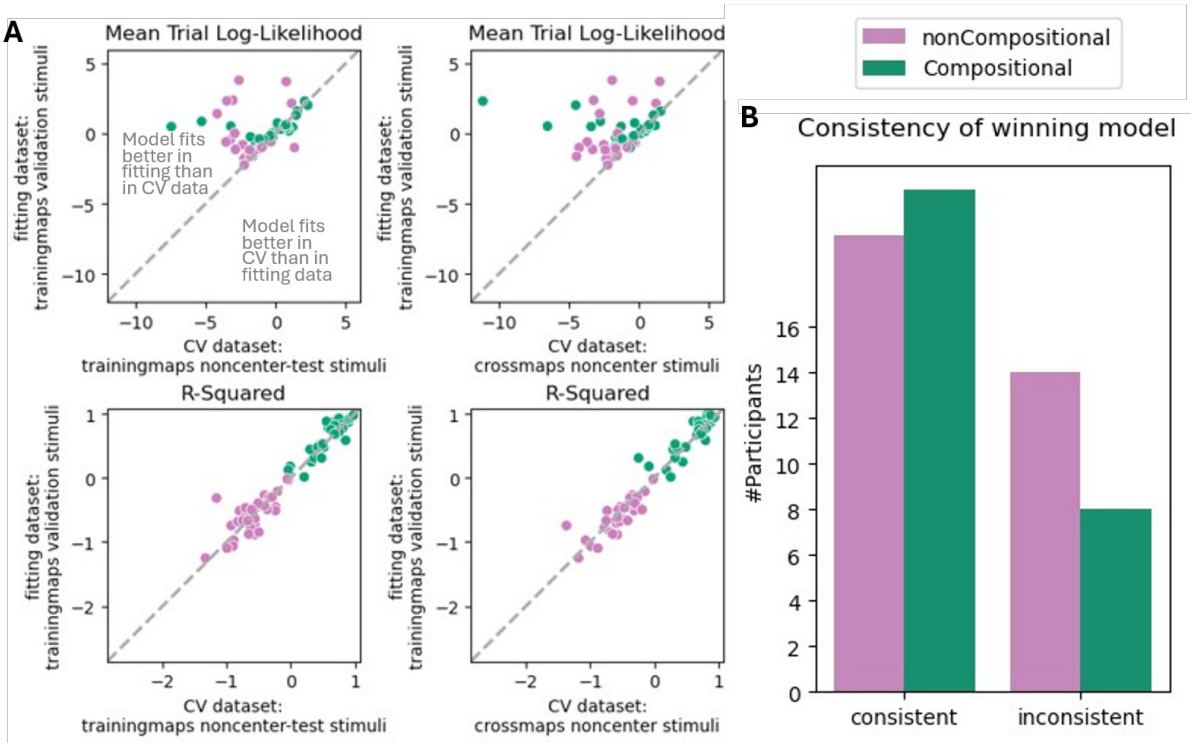
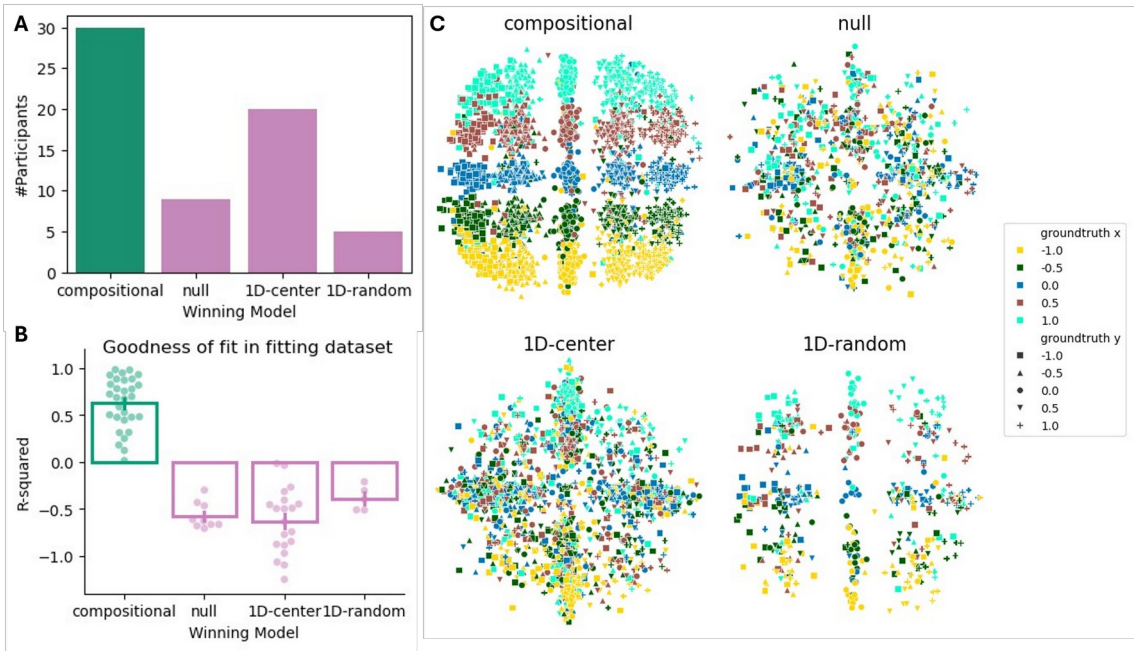


Figure 4-22. Winning model in cross-validation (CV) data. (A) Goodness of fit of the winning model in fitting data compared to its performance in different CV data subsets. Goodness of fit was measured by two metrics: mean trial log-likelihood (top panels) and R-squared between simulated data and observed data (bottom panels). Models were fitted using data from the non-centre validation stimuli in training maps (fitting data). Two CV data subsets were plotted (panels from left to right): non-centre-test stimuli on the training maps and non-centre stimuli on the cross maps. Coloured dots indicated goodness of fit for individual participants (purple: non-compositional, green: compositional). Dashed grey lines on the diagonal indicate that the winning model performed equally well in fitting and CV data. (B) Number of participants whose winning model was consistent in the fitting data and the CV data.

4.5.2.2 Goodness of fit of the winning models

Next, we examined the distribution of winning models among our participants and whether the winning model fit the response well. 30/64 of the participants were classified as compositional (Figure 4-9A). The remaining participants were distributed among the other three models instead of clustering in the null category. In fact, the largest group of non-compositional participants was the 1D-centre group (20/34 non-compositional participants). This suggested that the non-compositional participants were not just responding randomly. We further examined the goodness of fit of the winning models in the fitting data. When the compositional model was the winning model, the fitting quality was quite high (Figure 4-9B green bar, S.E.M.=0.625±0.114). However, the family of non-compositional models did not fit to the data as well as the compositional model (Figure 4-9B, R-square values in purple were below 0). This was also reflected on the response map of different groups of participants as classified by their winning models (Figure 4-9C). We can see that the



compositional participants' response map closely matched the simulated response in Figure 4-7, while this was not true in participants classified as 1D-centre and 1D-random.

Figure 4-23. Participant classification results. (A) Distribution of participants classified by their winning models. (B) Goodness of fit of the winning models in the fitting dataset. (C)Response maps of participants classified by their winning models. Each dot represented one participant's response to one stimulus in one map.

4.5.2.3 Behavioural Performance of compositional and non-compositional participants

The previous results established that the new classification method reliably identified *compositional* participants who solved the validation and test stimuli trials by combining rules learned in the supervised training trials. Therefore, we further investigated whether the behaviour patterns of the compositional and non-compositional participants in the current pilot experiment replicated those of the generalizers and non-generalizers from the previous pilots. We focused on two aspects: how well they learned and how they differed.

Similar to pilot M2V1, participants' performance approached the average performance in the fMRI study during pre-training (Figure S4-6). Participants on average missed 4.9% of trials in the 2AFC tasks, which was comparable to the previous pilot (Figure S4-2, Table S4-1). Taken together, we have solid evidence that participants can learn the compositional generalization rule in the multi-map task with stimuli generated from animal symbol combinations. Additionally, the transfer learning effect was replicated and confirmed by statistical analysis, whereby compositional participants learned T2 faster than T1 (Supplementary results 8.5, Figure S4-7).

The most important observation from the previous multi-map task pilot was that the training versus non-training stimuli dichotomy was no longer suitable. We found that the performance of generalizers and non-generalizers diverged more in trials with non-centre stimuli

than in trials with centre stimuli across maps. This was replicated in the treasure hunt task in the current pilot (Figure 4-10A, Supplementary Results 8.3). Finally, performance in the 2AFC task also replicated the results from the previous pilots. Participants' choice in the 2AFC task was consistent with their response in the treasure hunt task (Table S4-5, Figure S4-5, see Supplementary Results 8.5 for analysis method). The behavioural pattern in different participant groups and different stimulus groups also replicated the observations in the previous pilot (Figure 4-10B, see Table S4-4 for full mixed effect model results). In summary, compositional participants performed better than non-compositional participants (β=1.220, 95% C.I. [0.838, 1.602]). Accuracy was higher in trials with centre stimuli than in non-centre stimuli (β=0.332, 95% C.I. [0.270, 0.394]). And performance was sensitive to trial difficulty (β=0.054, 95% C.I. [0.020, 0.088]).

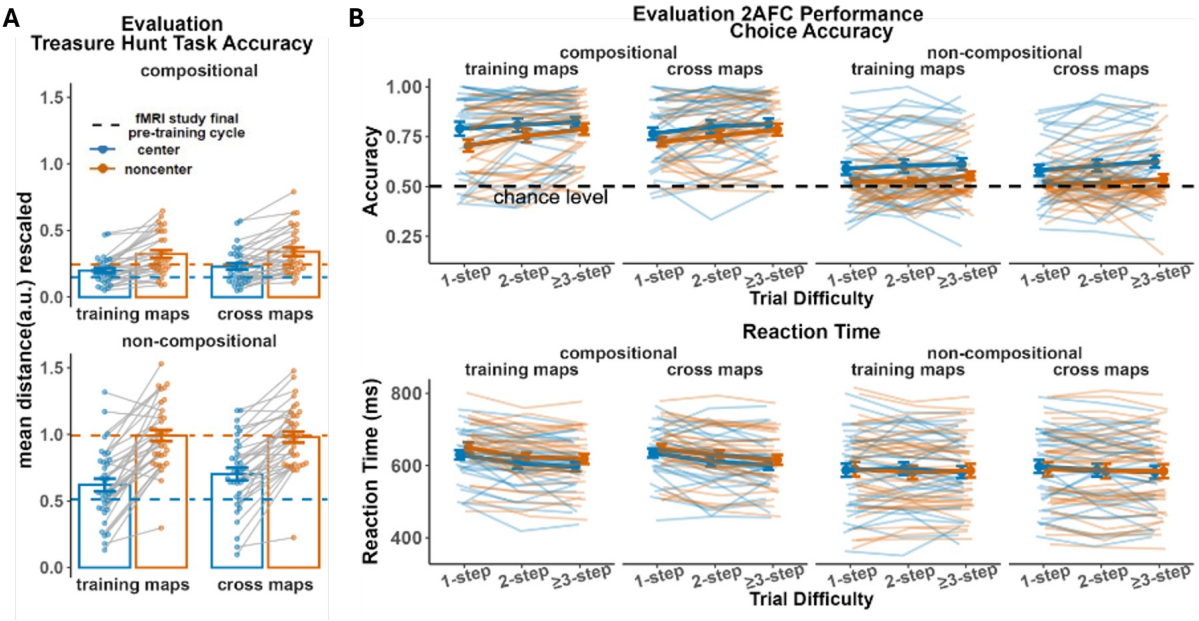


Figure 4-24. Participants' performance during the evaluation stage in two-map pilot M2V2-M2V3. Performance in (A) the treasure hunt task and (B) the 2AFC task is plotted. Dots and translucent lines represent individual participants. Bars and solid lines represent the group average.

Error bars represent mean ± standard error. Dashed lines in (A) represent group average performance in the one-map task from the fMRI chapter.

4.6. Symbol representation in the cluster task

We designed a symbol arena task to see if we could probe the internal representation of symbols for composing stimuli. At the end of the experiment, participants were asked to arrange the symbols to reflect the relationship between the symbols. This yielded a relational map for the symbols from each participant.

We were interested in the average relational map separately for compositional and non-compositional participants and wondered if it would reveal a parallel representation as in the fMRI study. To achieve this goal, we first computed the distance between symbols on this map. This yielded a behavioural representation dissimilarity matrix (RDM) for each participant. We rescaled each participant's RDM to the range of [0,1]. Then, we computed the average RDM of each participant group (Figure 4-11A) and performed multidimensional scaling (MDS) on the group average RDMs (Figure 4-11B) to visualise the average relational map for each group of participants. MDS revealed that compositional generalizers put the 20 symbols into two orthogonal axes, and the order of symbols matched their positions on the axes (Figure 4-11B, left). This structure was less clear in the non-compositional participants (Figure 4-11B, right).

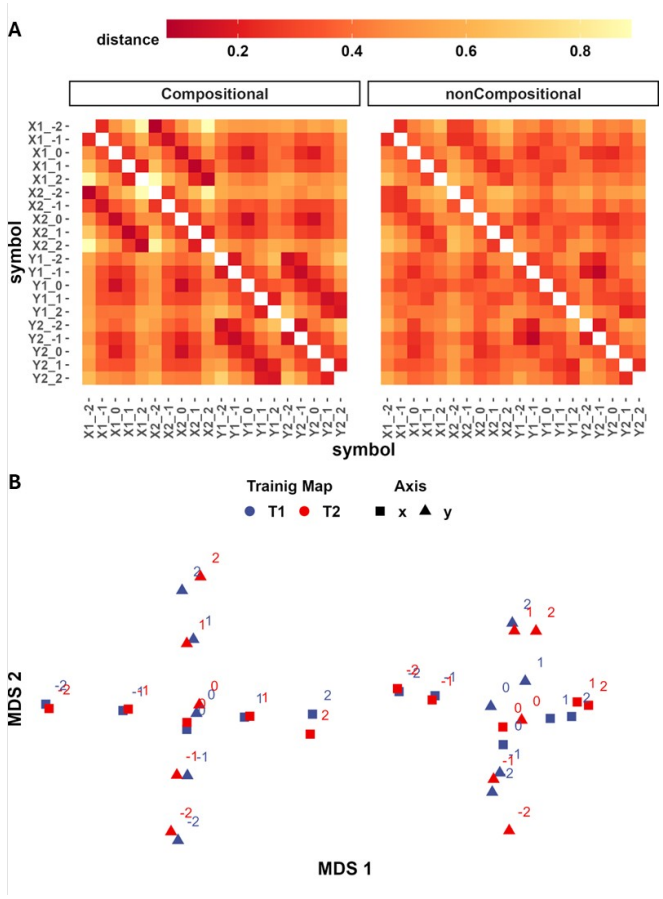


Figure 4-25. Symbol representation in the symbol arena task. (A) Average representation dissimilarity matrix (RDM) of compositional (left) and non-compositional (right) participants. (B) Multidimensional scaling plot of the group average RDMs. Text labels indicate the symbol locations on the axes (from -2 to 2).

4.7. Summary and Discussion

In this chapter, we designed a new version of the pirate task to address the limitation of the fMRI experiment in the previous chapter. We presented a series of pilot experiments that gradually refined the task design for a future fMRI study. Specifically, we introduced two major changes:

1) We generated stimuli with symbol-symbol combinations to make sure that any factorisation observed was not due to

inherently factorised sensory features (i.e., colour and shape in the previous fMRI chapter).

2) We increased the number of maps so that we can test if participants can combine rules learned in different map contexts.

To accommodate these changes, we developed a 2AFC task with a shorter trial duration to optimise for event-related fMRI. Our pilot experiments showed that our new design allowed us to address the limitations of the one-map task without sacrificing the behavioural accuracy:

1) In the new design, the proportion of participants who generalized compositionally and the average performance in the treasure hunt task were comparable to the one-map experiment reported in Chapter 3.
2) Participants showed flexible cross-context generalization in the multi-map task. they were not only able to combine the rules learned in the same training map, but also the rules learned in different training maps.
3) Importantly, the behaviour in the new 2AFC task was consistent with the behaviour in the original treasure hunt task.

Moreover, we developed a new classification method that disentangled failure to memorise the rules and failure to combine the rules for inference. In combination with the multi-map design, the new method allowed for cross-validation, which was not possible in the approach we used in the previous chapter. We showed that this method yielded consistent classification results in the majority of participants. Future work could further improve the classification consistency. In this chapter, we have obtained the maximum likelihood estimates (MLE) of the models' parameters using the L-

BFGS-B algorithm with warm starts. It is possible that the optimisation process obtained local minima; future work could employ multi-start grid search to find global minima. Given that all the models were mixture models, future work could also consider using the expectation-maximisation (EM) algorithm to obtain more stable parameter estimates more efficiently. Additionally, we observed that classification consistency and goodness of fit were worse in non-compositional generalizers. This may be due to less consistent behaviour among these participants. Alternatively, the issue could stem from our limited model space, which included only four models. A potential future direction would be to expand the classification analysis to the full model space.

Finally, we introduced a symbol arena task where participants freely arrange the symbols to see if we can probe participants' representation of the rules. On the positive side, compositional participants arranged the symbols in a way that reflected the symbols' locations on the axes. However, their arrangements of symbols from the x-axis and the y-axis were orthogonal to each other. This was different from the previous fMRI study, where we found that the hippocampus represented different axes as parallel manifolds. There were multiple possible explanations. It could be that participants relied on the spatial axes to organise the rule representation in the multi-map task, resulting in an orthogonal organisation of rules. Alternatively, the behavioural read-out of relational structures may be more similar to the representation in the parietal cortex instead of that in the hippocampus. However, it is hard to distinguish among these explanations purely from the behavioural data. Therefore, this task may not be suitable for testing the parallel representation hypothesis. Future work still needs to be done to find an appropriate behavioural marker for parallel representations.

Chapter 5 Transfer of structural knowledge

Contents

5.1 Introduction.....147
5.2 Experiment 1. Structure learning and transfer in human and neural network.....149
5.2.1 Method.....149
5.2.2 Results.....155
5.3 Experiment 2 Hidden layer representation that supports structure transfer in neural network.....162
5.3.1 Method.....162
5.3.2 Results.....163
5.4. Summary and Discussion.....170

5.1 Introduction

The world is not made up of events that are completely isolated from each other or haphazardly arranged. Instead, there are regularities and patterns, which make it possible to make predictions based on these relational structures. Take the family tree as an example, if we know that Alice and Bob are siblings, we can easily infer that their children are cousins. The ability to grasp and utilize relational structures is a hallmark of human intelligence. Psychometric instruments such as the Raven's Progressive Matrices Test (RPM) designed to test cognitive abilities in this respect have been shown to predict scholastic achievements and lifelong career success (Pind et al., 2003).

When we think in terms of relational structures, we go far beyond superficial physical properties (Tenenbaum et al., 2011). For instance, we can draw analogy between the trajectory of a planet and the trajectory of clock hands, because we identify the underlying cyclic structure even though clocks and planets are different. Studies on transfer, i.e., human's ability to repurpose learned knowledge to new settings, argue that induction of a general "problem schema" is required to facilitate successful transfer (Gick & Holyoak, 1983). Large body of literature show evidence that human can transfer learned knowledge within domain (*near transfer*) trivially but are less capable of transferring knowledge across domains (*far transfer*). A century ago, Woodworth and Thorndike (1901) reported that participants trained on estimating area of rectangles still perform poorly on estimating area of other shapes such as triangles, circles or irregular shapes even though these problems are closely allied. Similarly, Holyoak and Koh (1987) found that participants had a hard time spontaneously generalizing the learned solution when the component of the new problem is no longer superficially similar to the old one. These studies provide descriptive evidence of human's limited ability to perform far transfer. However, a formalized mechanistic account of how structure transfer can be achieved is yet to be established.

Equally important is the question of how structure knowledge is acquired in the first place. Educational practices make great endeavours to explore more effective curriculum to facilitate skill learning and transfer. A common rule-of-thumb principle is to start small. Unlike machines that can process high-dimensional information in parallel, humans have limited cognitive capacity and tend to rely on temporal statistics to identify the structure underlying a stream of information. Previous research has found that participants can successfully separate task sets with different contingency rules even

without explicit cues (Collins & Koechlin, 2012; Collins, 2017). In the inductive categorical learning literature, behavioural and eye-tracking studies revealed benefits of training sequences that control for the confounding factors and highlight diagnostic features of each category (Abel et al., 2021; Carvalho & Goldstone, 2015, 2017; Zaki & Salmi, 2019). Understanding human's sensitivity to temporal statistics not only informs curricula design but also provides a window to uncover learning principles in the biological brain. On a computational level, it hints that a temporal asymmetric learning mechanism could give rise to the "slow learning" of high-order structures (Botvinick et al., 2019; Kumaran et al., 2016). In object recognition, it has been shown that the extraction of slow-varying features leads to learning transform-invariant representations of objects (Rolls, 2021). While in more complex tasks, chunking of temporally contiguous events driven by the Hebbian learning rule subserve the autonomous learning of task sets (Bouchacourt et al., 2020) and temporally autocorrelated training regimes benefit the formation of factorized representations in multi-task settings (Flesch et al., 2018) and compositional generalization (Dekker et al., 2022).

With both questions in mind, the current project aims to elucidate the acquisition and generalization of structure knowledge. Inspired by the pirate task that was used to investigate compositional rule generalization in the previous chapters, we develop a paradigm in which different stimulus features were embedded with distinct relational rules that collectively determined the mapping from stimuli to labels. Crucially, we train participants on the mapping using one set of labels and test whether they can generalize the underlying relational structure to correctly predict the mapping from the same set of stimuli to a new set of labels — much like a person transferring their knowledge of how to use a coffee machine with English button

labels to one with labels in Chinese. Furthermore, we aim to test whether the induction of structural form in humans also operates on the temporal statistics of the environment. We operationalize this by introducing blocking, a temporal sequence of stimulus presentation in which one feature dimension varies slowly over time. We predict that this manipulation will facilitate structural learning and generalization in humans. Lastly, using connectionist models, our aim to characterize the representational format that supports structure learning and transfer.

5.2 Experiment 1. Structure learning and transfer in human and neural network

5.2.1 Method

5.2.1.1 Task

Our task asks if participants can learn the covariance structure among variables. We construct 27 stimuli from the factorial combinations of three colours, three animal shapes, and three textures (Figure 5-1A left panel). Participants are asked to predict labels (pseudowords) for these visual stimuli. Unbeknownst to the participants, there exists a relational pattern: blue and yellow stimuli of the same shape always have the same label (Figure 5-1A right panel). For example, the label of a blue whale is the same as a yellow whale, a blue elephant is the same as a yellow elephant. The fact that blue and yellow always go together makes colour relevant to the relational structure, thus, we call it the relevant dimension. The “blue is the same as yellow” relational structure is preserved across different shape instances, but stimuli of different shapes have different labels. That is, shape adds constraints to the relational structure, so we refer to shape as the constraining dimension. Lastly,

stimuli of the same colour and shape share the same label regardless of texture, which makes texture an irrelevant dimension. Based on this structure, all 27 stimuli can be divided into 6 clusters and stimuli in one cluster have the same label (Figure 5-1A right panel). Note that the mapping between feature and dimension is counterbalanced across different participants (Table S5-1A). If we disregard the irrelevant features, the stimuli-label mappings in the task can be visualized using a 3×3 table (Figure 5-1B), so we refer to this version as *the 9-grid task*.

We hope to quantify two aspects of behaviour in this task: 1) how well can participants learn the stimuli-label mapping from feedback? 2) how well can participants learn the structure knowledge and generalize it to a new response space? To this end, we ask participants complete 10 training blocks and 10 transfer blocks in an interleaved order (i.e., 10 train-transfer cycles). During training blocks, participants received full feedback on the association between 27 stimuli and one set of 6 distinct labels. To test if participants can generalize the structure knowledge acquired in training, we ask participants to map the same 27 stimuli to a new set of 6 labels in transfer blocks. Because the mapping from 6 stimuli clusters to 6 new labels is not unique, we provide additional constraints (“anchors”) at the beginning of transfer blocks so that the problem only has one solution. At the beginning of transfer blocks, we randomly present participants with 15 stimuli that are associated with 5 labels in the second set with feedback (Figure 5-1B). We refer to these feedback trials in transfer blocks as *anchor trials*. After that, test trials are presented in random order and without feedback. Therefore, to effectively perform the task, participants need to identify the underlying structure during training and generalize it to a new set of labels.

Note that there is always one stimuli category, i.e., one label, that participants do not see in anchor trials. Hence, participants need to rely on counterfactual inference (“if not A then B”) to correctly predict the label of this category in test trials. Across participants, this unseen stimulus category in anchor is randomly chosen either from one of the correlated pairs (denoted as C, blue or yellow stimuli categories in Figure 5-1B) or from the standalone ones (denoted as U, red stimuli categories in Figure 5-1B).

Each stimulus is presented once in each block, so each block consists of 27 trials. In each trial, one stimulus is presented to participants after 1 second of fixation and all 6 labels in the current block are presented to participants on a ring surrounding the stimulus (Figure 5-1C). Participants are required to choose the label associated with the current stimulus. Responses are self-paced. In training trials and anchor trials, participants receive feedback on their responses (Figure 5-1C). No feedback is provided in test trials. For each participant, two sets of labels are used, and all 12 labels have different initials. Labels are sampled from the NOUN dataset (Horst & Hout, 2016) using a weighted random sampler (the weight of a given initial is defined as the number of pseudowords that start with that initial).

5.2.1.2 Design

Another goal of the current project is to test how different training curricula affect participants’ learning and generalization performance. We adopt a between-group design. During training, we introduce temporal autocorrelation (blocking) to one of the three dimensions, i.e., the constraining dimension, the relevant dimension, or the irrelevant dimension. More specifically, each training block can be divided into 3 mini-blocks (no interval between mini-blocks). Each mini-block randomly presents 9 stimuli that have the same constraining feature, relevant feature, or irrelevant feature. That is,

we split the 3-d stimulus space (Figure 5-1A) into three slices along the blocked dimension, and each mini-block randomly presents stimuli in one slice. For example, if shape is the constraining dimension and is blocked, participants would first see 9 whales, and then 9 ducks, and finally 9 elephants in a training block. Based on our hypothesis that blocking features that define the covariance structure benefit learning, the first two groups (i.e., blocking the constraining or relevant dimension) are our treatment groups, while the third group (i.e., blocking the irrelevant dimension) is the control group.

5.2.1.3 Participants

Data is collected online with participants recruited from Prolific. They were randomly assigned to one of the three conditions (constraining, relevant, or irrelevant condition) and one of the six feature-dimension mappings. Each block has an attention check trial at the end in which participants have to choose the novel label unseen in this block. 4 participants who fail attention checks in more than three blocks are excluded. 5 participants who report that they took notes during the experiment are excluded. The following analysis is based on data from 90 participants (28 in the constraining condition, 33 in the relevant condition, and 29 in the irrelevant condition, Table S5-1B).

5.2.1.4 Neural Network Simulations

We build an artificial agent to perform the 9-grid task using PyTorch. It is a feedforward neural network with one hidden layer of the form

$$y = W_2 * ReLU\left(W_1 * x\right)$$

Each feature (colour, shape, texture) is one-hot encoded by the input vector x. Input vector x also has two additional context units as one-hot coding for the label set available. Therefore, input vectors have 11 entries for the 9-grid task. The input layer is fully connected to the hidden layer with a weight matrix W_1. The rectified linear unit function

is used as the hidden layer activation function to introduce nonlinearity.

$$ReLU(z)=max(0,z)$$

The hidden layer is then fully connected to a single output layer with a weight matrix W_2. The output layer is a one-hot encoding of all the labels in the task. Hence there are 12 output units for the 9-grid task. The behaviour of this artificial agent is fully deterministic using the argmax rule, i.e. the label that corresponds to the output unit with the largest value y is chosen as the response for the current trial.

We train the network using stochastic gradient descent and cross-entropy loss. The implementation in PyTorch is equivalent to the combination of log-softmax and negative log-likelihood loss.

$$loss(y,t)=-\sum_{i=1}^{12}\log\frac{\exp(y_i)}{\sum_{j=1}^{12}\exp(y_j)}*y_i$$

t is the target vector (supervision signal), y is the output vector, y_i is the output value of the ith output unit.

In each epoch, the agent is exposed to one training-transfer cycle as human participants do. Supervision signal is only provided during training trials and anchor trials in transfer. Network parameters are updated after each supervised trial. The length of training varies according to hidden layer size.

We collect simulation data on both tasks with multiple runs, each with unique random initializations. We run simulations with varying hidden layer sizes (4,5,6,7,8,9,10, and 20 hidden units). 30 runs are collected for each hidden layer size. In each run, models with the same initialization received full training of different curricula. To further illustrate that the training on structure knowledge is vital for

generalization performance, for each curriculum, we set up two control training regimes in addition to the full training regime, namely the “anchor only” and the “train random” regime. In the “anchor only” regime, we take away the training block in each epoch, and the agent only received supervision signals during anchor trials in transfer blocks. In the “train random” regime, stimuli-label mapping of the training block is randomly permutated before the experiment, hence breaking the covariance structure. Each run has a unique random permutation. Then the agent undergoes the same anchor training and test procedure as described above. Coherent with the human experiment, the unseen stimuli category in anchor trials can be either C or U (Figure 5-1B). In brief, for one random initialization, we run 18 variants of simulations varying training curriculum, training regime, and counterfactual inference condition (the unseen category in anchor).

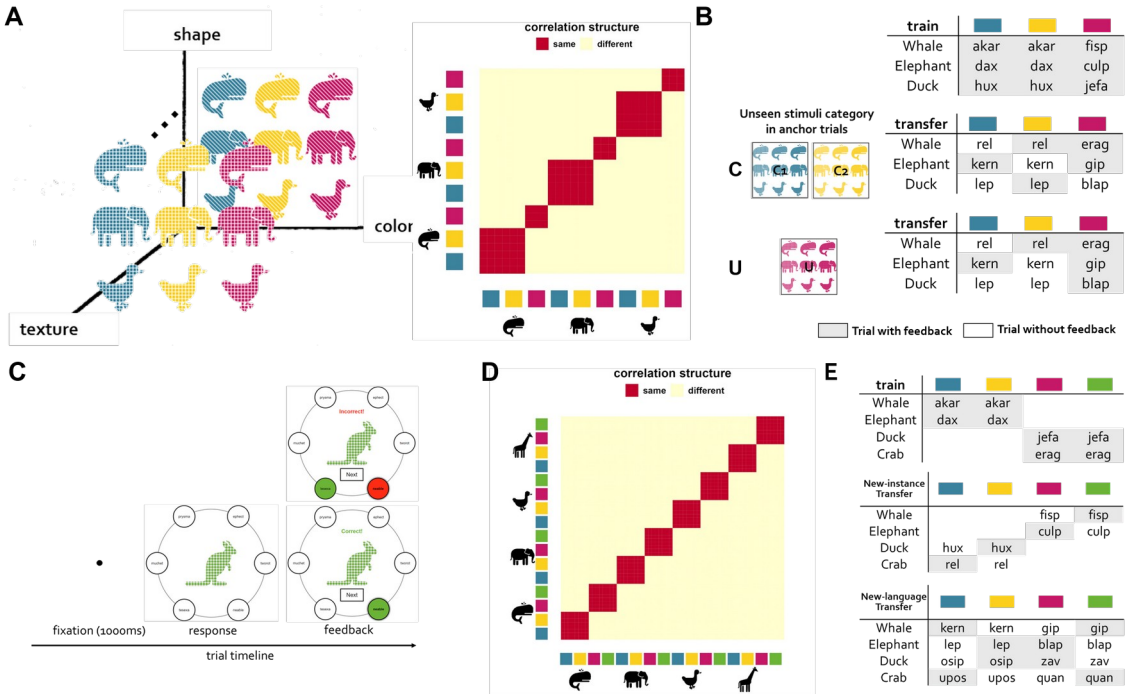


Figure 5-26. Task Design. (A) Stimuli features and the relational structure of the 9-grid task. Left panel: Illustration of the three-dimensional stimuli

space. Right panel: Illustration of the relational structure using a correlation matrix. The Colour of each grid indicates whether the stimuli pair has the same label. (B) Illustration of stimuli-label mappings in training and transfer blocks in the 9-grid task. Grids with grey backgrounds represent trials with feedback (i.e., all trials in training, and anchor trials in transfer). Grids with white backgrounds represent trials without feedback (i.e., test trials in transfer). One set of labels is used in training and a second set is used in transfer, but the structure of stimuli-label mappings is preserved across blocks. One stimuli category is not presented in anchor trials. Across participants, this unseen stimuli category is randomly chosen from the correlated pairs (middle blue or yellow stimuli) or the standalone ones (red stimuli). (C) Illustrated here is the timeline for training trials and anchor trials. The timeline for test trials is similar but without the feedback in the end. (D) The relational structure of the 16-grid task. The colour of each grid indicates whether the stimuli pair has the same label. (E) Same as B but for the 16-grid task. Only grids with labels are included in the corresponding train/transfer block.

5.2.2 Results

5.2.2.1 Participants learn and generalize the covariance structure and benefit from certain training curricula

In our 9-grid task, human participants learn the relational structure through repetitive training and generalize this knowledge to make novel inferences in transfer (Figure 5-2). We observe that performance in training is way above chance from the very first block (One-sample t-test on the first training block's accuracy compared to a chance level of 1/6, constraining condition: $t(27) = 10.642$, constraining condition: $t(27) = 10.642$, relevant condition: $t(27) = 9.024$, irrelevant condition: $t(27) = 3.700$, $p<0.001$ for all conditions and all p values are FDR-corrected) and increases over time (Table S5-2A). As a new set of labels is used in transfer and no feedback is shown in test trials, performance in test trials can be used as an index of structure learning and transfer. Participants' performance in these

test trials is significantly above chance from the second transfer block (Table S5-2D).

Blocking the constraining or relevant dimension benefits structure learning and transfer (Figure 5-2). Compared to the control group (i.e., blocking irrelevant dimension), both treatment groups (i.e., blocking constraining or relevant dimension) achieve higher overall accuracy and improve faster during training (Table S5-2A), but there is no difference between the two treatment groups (Table S5-2B). In addition, generalization performance in test trials of both treatment groups increase faster over blocks compared to the control group (Table S5-2A). Again, there is no significant difference between the two treatment groups (Table S5-2C).

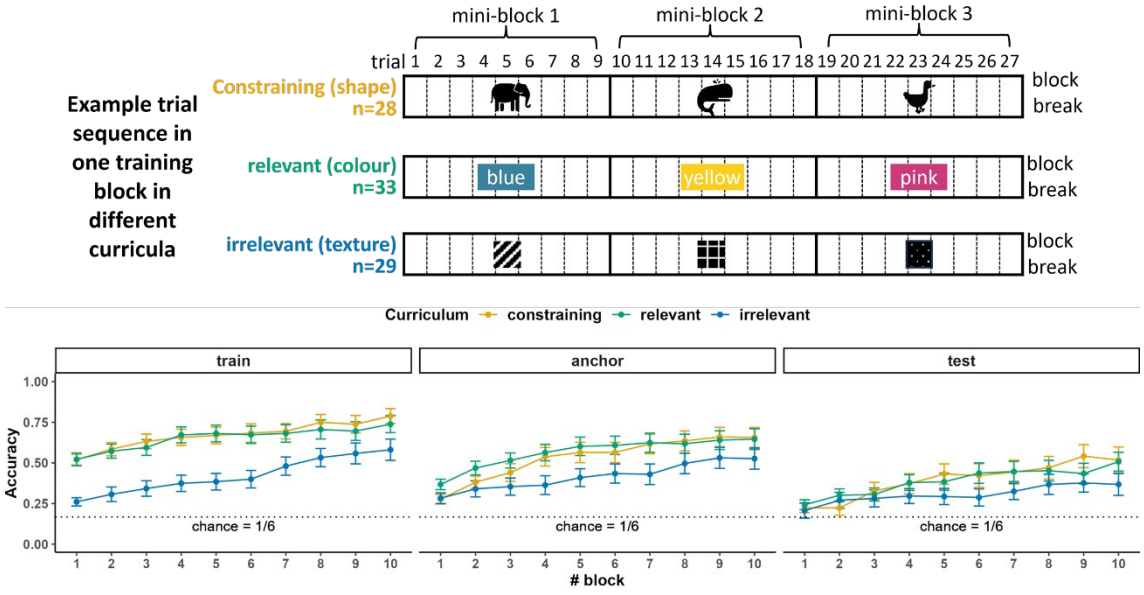


Figure 5-27. Structure learning and transfer performance in the 9-grid task in humans. Top: example trial sequence of different training curriculum. Bottom: Participants' accuracy as a function of block number, trial type, and condition. The dashed black line indicates a chance level of 1/6. Error bars indicate mean± standard error.

5.2.2.2 Blocking constraining and relevant dimension lead to different error patterns

Next, we probe the difference between training regimes in the treatment groups. We build representational dissimilarity matrices (RDM) to represent the structure of different models as well as the structure that participants have learned. Each RDM is a 27*27 matrix, the value of each grid represents whether the corresponding stimuli pair share the same label. We derive four RDMs to model four possible structures, one ground-truth structure, and three approximate structures (Figure 5-3A). For participants' behaviour, we compute an RDM for each participant and in each block. Group-level RDMs are computed by averaging participants' RDM in each condition. As full feedback is provided during training, the similarity between observed participants' RDM and model RDMs can be driven by the acquisition of stimuli-label associations rather than structure learning. Thus, we focus on the structure of participants' responses during transfer.

Training curricula impacted the emergence of different structures in participants' RDMs (Table S5-3A, Omnibus test of ground-truth RDM * time * condition interaction, $\chi^2(2) = 166.503$, $p<0.001$, FDR-corrected). The ground-truth structure grows faster across blocks in the training regime that blocks constraining or relevant dimension than the one blocking irrelevant dimension (Figure 5-3B, Table S5-3A). Moreover, post-hoc comparisons reveal that it is also faster in the constraining condition than the relevant condition (Figure 5-3B, Contrast Estimate = 0.018±-0.002, $Z = 9.225$, $p < 0.001$, FDR-corrected).

To compare the contribution of different model RDMs, we regress group-level RDMs on different model RDMs separately for each block and each condition and tested whether the coefficient of the ground-truth RDM regressor is significantly different from that of the constraining RDM regressor. Results showed that different structures

are induced in different training regimes in the early stage of the experiment (Figure 5-3B). In the first three transfer blocks, the blocking-relevant group showed greater similarity with the constraining model RDM, and then gradually converge to the ground truth RDM after block 4 (Table S5-3B). This pattern is not observed in the other two groups.

The above analysis suffers from the issue of collinearity between different model RDM regressors. To further disambiguate, we focus on the different error predictions of these models. Consider a counterbalance instance in which shape is the constraining dimension, colour is the relevant dimension, and textures are irrelevant (i.e., blue and yellow stimuli of the same shape have the same label). We define two types of errors based on the current stimuli category and the stimuli category that the mistakenly chosen label belongs to (Figure 5-3C). If a participant chooses the label that corresponds to a stimulus that has the same shape as the current stimulus but with a different colour, this is a colour error. Likewise, a choice that would be otherwise correct for a stimulus of the same colour but a different shape than the current stimulus is classified as a shape error. Hence, the constraining model RDM makes colour errors as it predicts that stimuli of the same shape share the same label. Similarly, the relevant model RDM makes shape errors. Errors that do not fall into these categories are called other errors.

We explore the error pattern in participants' data using a three-way mixed ANOVA on participants' average percentage of errors aggregated over all blocks (Figure 5-3E). As we are only interested in the two contextual errors, trials classified as other errors are excluded from this analysis. There is a significant three-way interaction between condition, block type and error type ($F(2,87) = 4.3$, $p = 0.017$). Post-hoc analysis reveals that the interaction between

condition and error type is only significant in transfer ($F(2,87) = 8.39$, $p = 0.003$, FDR-corrected), but not in training($F(2,87) = 0.235$, $p = 1$, FDR-corrected). Further breaking down the effect, blocking shapes (the constraining condition) during training leads to more colour errors than shape errors in transfer (Wilcoxon signed-rank test on paired samples, $W = 99$, $p = 0.017$, FDR-corrected). On the contrary, blocking colour (the relevant condition) during training leads to more colour errors than shape errors in transfer (Wilcoxon signed-rank test on paired samples, $W = 415$, $p = 0.015$, FDR-corrected). This result holds if all error trials are included in the analysis, or if we only focus on the test trials during transfer.

The observed error pattern can be explained if we think of these two error types as contextual errors that arise from the context provided by different training curricula (Figure 5-3D). When shapes are blocked (i.e., in the constraining condition), participants first see nine elephants of different colours and textures, then nine ducks, and finally nine whales. In this curriculum, shapes become contexts and lie on a higher hierarchy than colours. Participants are exposed to one context at a time and receive extensive training on the difference between instances in that context. Thus, they are less likely to be confused between instances under the same context than to be confused between different contexts (Collins & Koechlin, 2012). Therefore, participants in the constraining condition are more likely to make shape errors than colour errors. Similarly, participants in the relevant condition would use colours as contexts and make more colour errors.

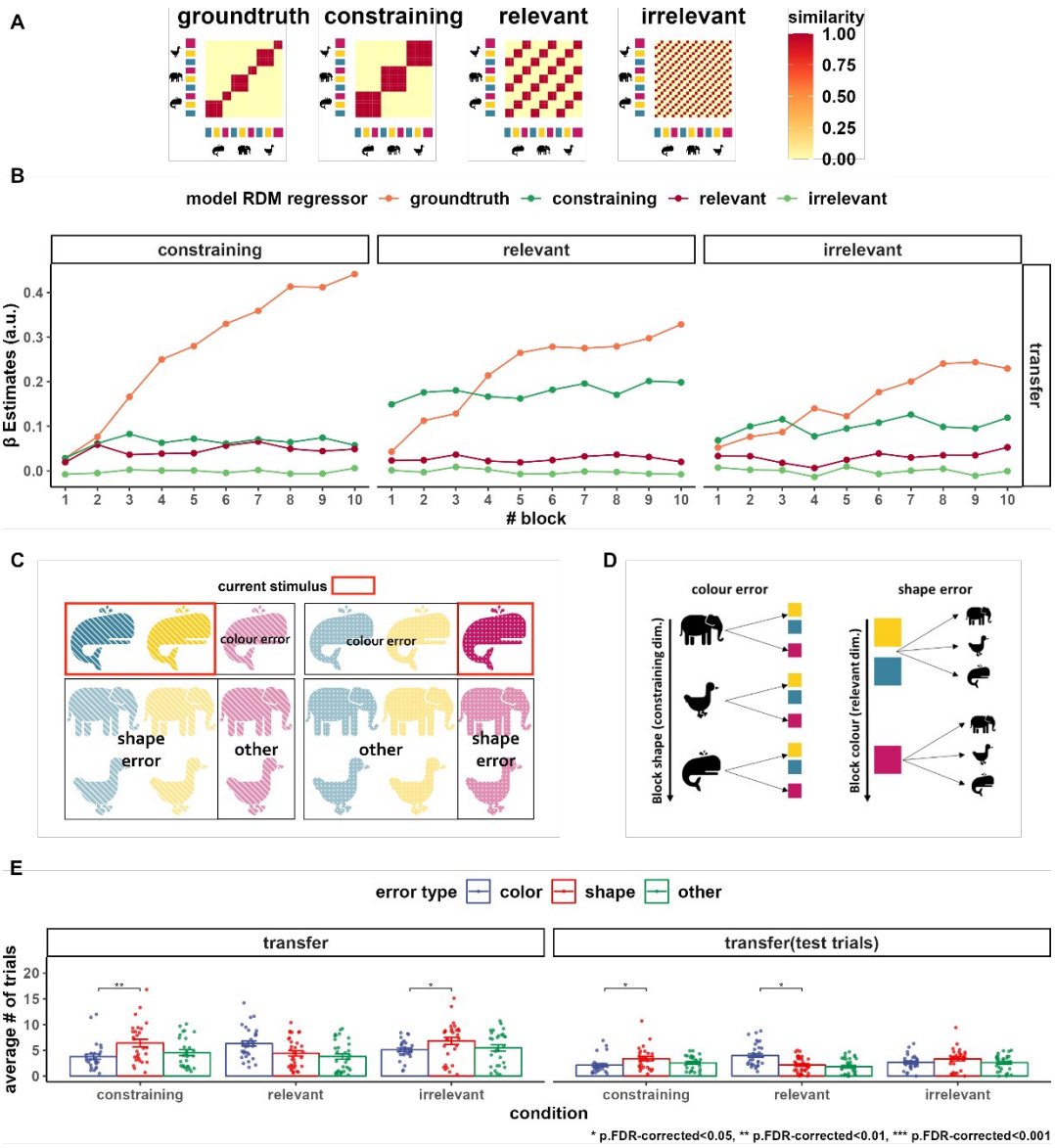


Figure 5-28. Effect of training curriculum on participants' learning and generalization behaviour. (A) visualization of model RDMs. Groundtruth model RDM: the true correlation matrix of the experiment design; Constraining RDM: stimuli that have the same constraining feature (e.g., shape) share the same label; Relevant model RDM: stimuli that have the same relevant feature (e.g., colour) share the same label; Irrelevant model RDM: stimuli that have the same irrelevant feature (e.g., texture) share the same label. (B) Coefficient estimates of different model RDM regressors when group-level RDM of different blocks and different conditions are fitted separately. Each panel shows the comparison between RDM regressors in one training curriculum group. (C) Error classification. The red box indicates current stimuli. Black boxes indicate error types. (D) Illustration of how different curricula may lead to different types of

confusion. (E) The percentage of errors as a function of error type, block type, and condition. Error bars indicate mean ± standard error. Asterisks indicate the results of pair-wise comparisons using paired Wilcoxon tests.

5.2.2.3 Vanilla neural network fails to make counterfactual inferences and does not benefit from curricula

We train a simple feedforward neural network with one hidden layer to perform the same task as humans do. Network converges to 100% accuracy in training trials and anchor trials if it undergoes full training of different curricula. Compared with the full training regime, the "anchor only" and "train random" regimes that take away training on the relational structure are detrimental to performance in the test trials during transfer, but there is no curriculum effect (Figure 5-4A). However, unlike humans (Figure 5-4C), this vanilla neural network architecture cannot make counterfactual inferences. In test trials, the counterfactual inference is required when stimuli are from the category unseen in anchor trials (Figure 5-1B, 6/12 trials if the unseen stimuli category is from the correlated pairs (C), 3/12 trials if it is one of the uncorrelated standalone categories (U)). While there are human participants who can solve this and continue to improve throughout the experiment (Figure 5-4C), the network fails to get any of these trials correct (Figure 5-4B). Therefore, network test performance plateaued at an accuracy of either 50% or 75% depending on which stimuli category it hasn't seen during anchor trials (Figure 5-4A, first and fourth rows right panels). Increasing the hidden unit size from 4 to 20 didn't change these results.

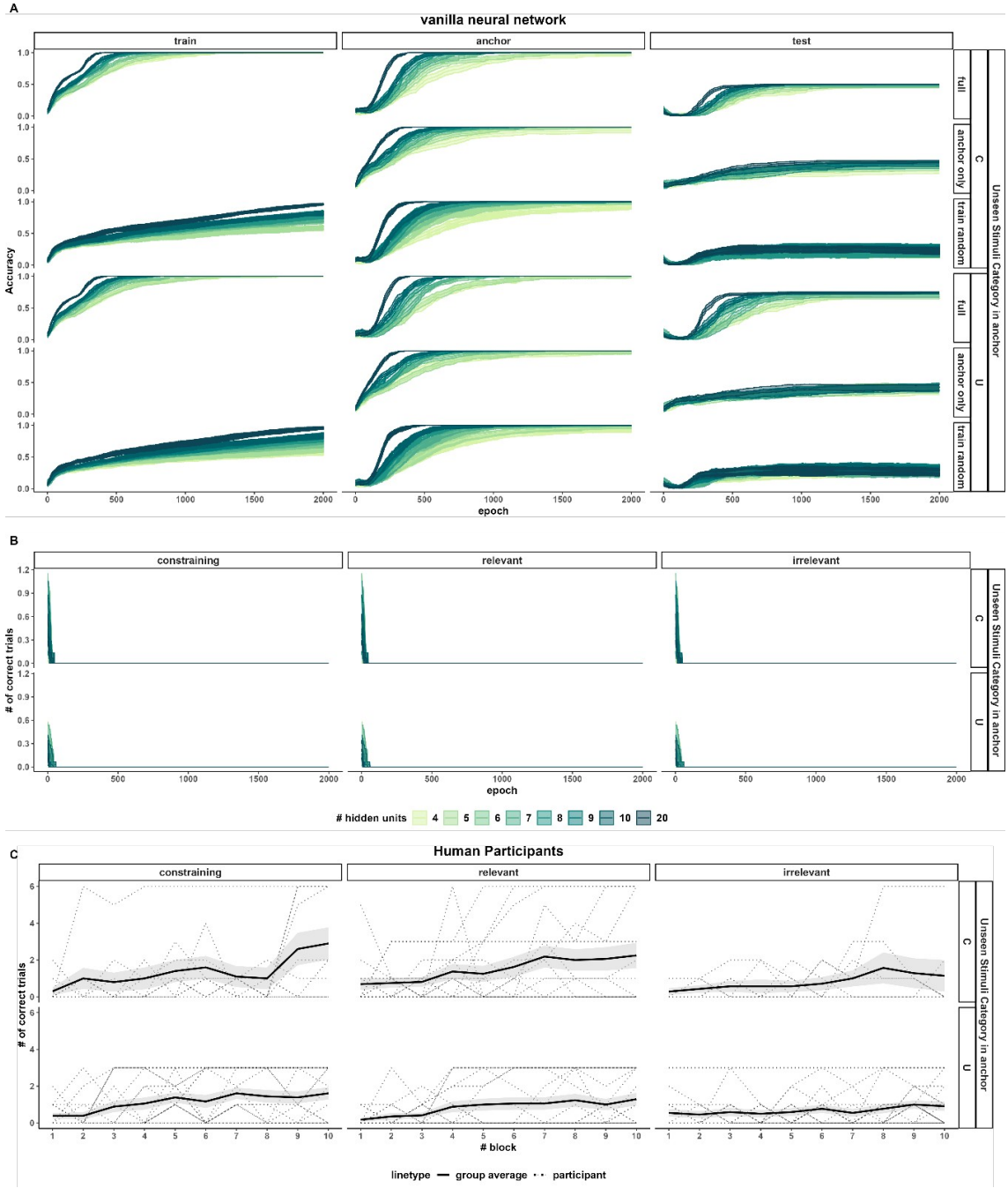


Figure 5-29. Structure learning and transfer performance in the 9-grid task in humans. (A) Neural network performance in train (left panels), anchor (middle panels) and test trials as a function of epoch. Accuracies are plotted in separate rows for different counterfactual inference conditions (unseen stimuli category) and different training regimes. In each facet, colour indicates the number of hidden units in the network, while line type indicates curriculum. (B) Neural network performance in test trials of different counterfactual inference conditions (unseen stimuli category) (C) Same as B but for human participants. Lines indicate mean. Shaded ribbons indicate mean± standard error.

5.3 Experiment 2 Hidden layer representation that supports structure transfer in neural network

As counterfactual inference is not the main focus of the current project, we develop a new variant of the label prediction task with a covariance structure similar to the 9-grid task. We run a new simulation experiment on this task with the same network architecture as in Experiment 1. We expect the network to solve the task perfectly when counterfactual inference is not required. If so, we can further look at the hidden layer representation that supports structure transfer.

5.3.1 Method

5.3.1.1 Task

We eliminate the counterfactual inference component of the 9-grid task and expand the task into a 4×4×4 feature space. We refer to this version as *the 16-grid task* hereafter. In the 16-grid task, there are two pairs of correlated colours: blue-yellow and red-green (Figure 5-1D). During training, the agent is exposed to 4 grids in the top-left corner and 4 grids in the bottom-right corner (Figure 5-1E upper table). This partial training setup allows us to examine structure transfer in two settings. The first setting aims to test if agents can transfer the learned colour-pairing rule to new shape instances (referred to as *new-instance transfer* hereafter). For example, if they learned that a blue whale is the same as a yellow whale, can they predict that a blue duck is the same as a yellow duck? Agents receive information about stimuli on the left diagonal in anchor trials and are tested on the correlated off-diagonal ones (Figure 5-1E middle table). The second transfer setting is the same as the transfer block in the 9-grid task. A new set of labels is introduced, but the covariance

structure is preserved. It is as if agents are transferring structure knowledge to a new language. Agents receive information about stimuli on both diagonals in anchor trials and are tested on the off-diagonal ones (Figure 5-1E bottom table).

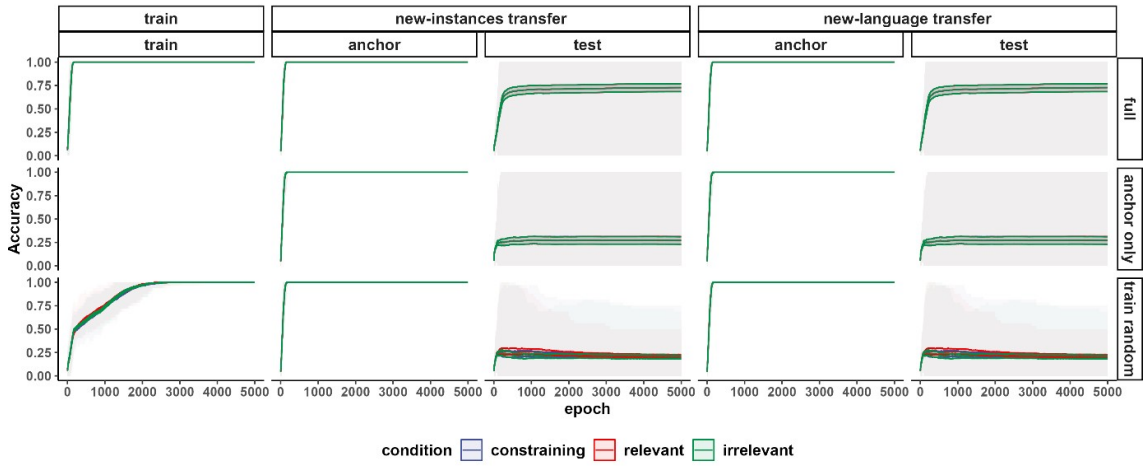
5.3.1.2 Neural Network Simulations

We use the same architecture as in experiment 1 with adjustments on the number of input, output and hidden layer units. The input vector has 18 entries and the output vector has 16 entries. The hidden layer has 20 units. The simulation procedure is also identical to experiment 1. Note that for the 16-grid task, each epoch is a train-transfer cycle which consists of training, new-instance transfer and new-language transfer. The presentation order of the two transfer settings is randomized on a per-run basis. We collect data from 60 runs to make sure we have enough data from both orders. Similar to experiment 1, for one random initialization, we run 9 variants of simulations varying training curriculum and training regime.

5.3.2 Results

5.3.2.1 Vanilla neural network performs the 16-grid task to ceiling

Consistent with findings in experiment 1, training curricula do not influence the performance of this vanilla neural network training regime, and lack of pre-training on the relational structure is detrimental to performance in the test trials during transfer (Figure 5-5, second and third rows). Intriguingly, when averaged across all runs, the mean test performance in the full training regime didn’t reach 100% (Figure 5-5, first row). In 45% of independent runs, the network could reach 100% accuracy in test trials during both new-instances transfer and new-language transfer settings as predicted. There exists great variability in test performance across runs even though



performance in training and anchor trials converged in all runs. This provides a testbed to examine the hidden layer representation that supports transfer in test trials.

Figure 5-30. Structure learning and transfer performance in the 16-grid task in the neural network. Training and transfer performance of the neural network as a function of epoch number, trial type, condition, and training regime. Solid lines indicate mean accuracy, while ribbons indicate mean± standard error. Shaded areas in grey indicate the range of accuracy across all 60 independent runs.

5.3.2.2 A theoretical representation geometry that supports structure transfer to new instances

The 16-grid task essentially requires the network to learn to put all 64 stimuli into 8 clusters during training and find the mapping between labels and clusters during test trials based on information provided in anchor trials during transfer. We ask what would be a useful representation of these 8 clusters that would benefit generalizing the learned colour pairing rule to new shape instances.

A high-dimensional representation of all eight categories would be a regular 7-simplex formed by 8 vertices equidistant from each other in a seven-dimensional space (Figure 5-6A left panel). In this model, the between-cluster similarity is always 0 and information cannot be shared between clusters (Figure 5-6A right panel). As a result, this will

hinder the generalization of the relational structure between blue and yellow from whales to elephants.

Previous literature suggests that abstract representation that reproduces the latent structure in the sensory world supports knowledge generalization (Bernardi et al., 2020; Johnston & Fusi, 2022). Based on these findings, we design a model of ideal representation geometry that encodes the two variables that define the covariance structure in an abstract format: correlated colour pair and shape.

Consider a linear classifier trained to separate the two correlated colour pairs in elephant stimuli. To perfectly generalize to whale stimuli, the hyperplane separating colour pairs when shape is elephant should be parallel to its counterpart when shape is whale (Figure 5-6B). In other words, the coding direction of colour pairs under different shape instances should be parallel with each other. If these hyperplanes are orthogonal, the agent would fail to distinguish between different colour pairs and make colour errors. Likewise, the coding direction of any shape dichotomy (e.g. elephant and whale) should be parallel across colour pairs (Figure 5-6B).

Taken together, we derive the parallel tetrahedron model (Figure 5-6C left panel). Within each correlated colour pair, shapes are represented as the four vertices on a regular tetrahedron, meaning that four clusters within a colour pair are equidistant from each other. Assuming homogeneity of shape representation across colour pairs, the tetrahedrons for the two correlated colour pairs should be identical. Similarly, assuming homogeneity of colour pair representation across shapes, the tetrahedrons should be parallel. We make the additional assumption that there is no shape or colour bias in the representation, which means that the distance between the

two identical tetrahedrons equals the edge length of each tetrahedron.

We predict that the performance variability can be attributed to different hidden layer geometry. That is, hidden layer representation should resemble the parallel tetrahedron model after full training, especially in the runs where the test accuracy is high. If we look at the geometry from a sample run that reaches 100% accuracy in test, it indeed looks like the ideal parallel tetrahedron model (Figure 5-6D).

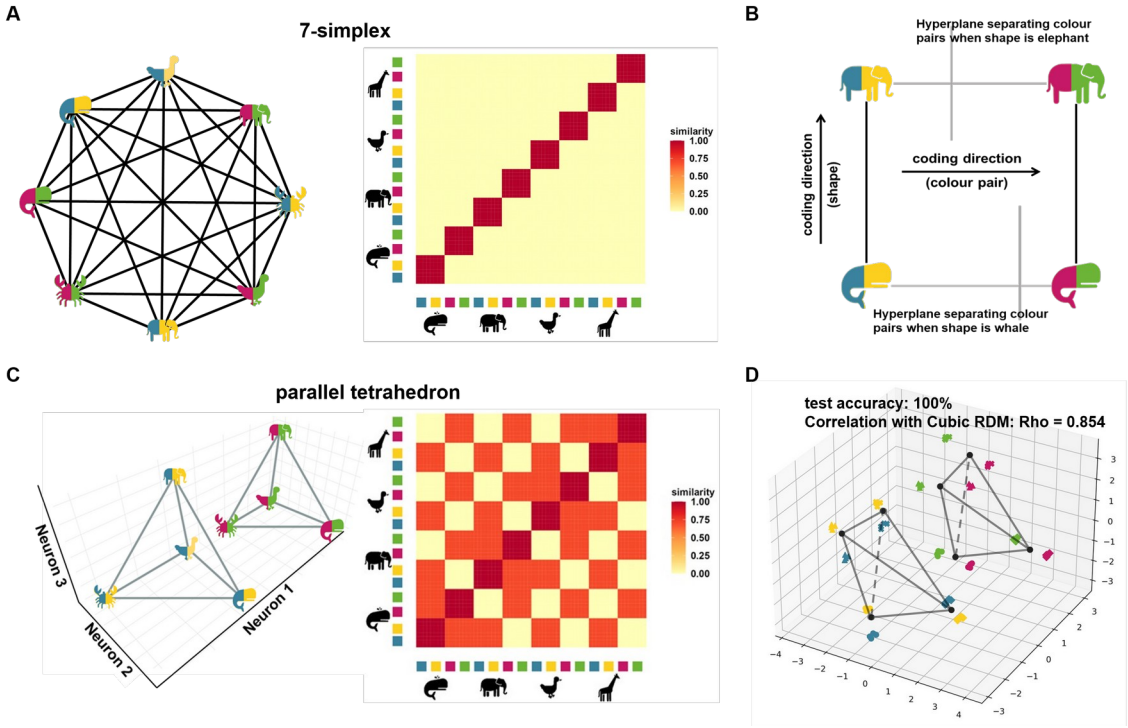


Figure 5-31. Proposed representational geometry that supports structure transfer. (A) Representational similarity matrix of a manifold formed by 8 equidistant points lying in seven-dimensional space. (B) A cubic model of representational geometry that supports structure transfer. (C) Same as A but for the cubic model. (D) Multidimensional scaling of the hidden layer representation of training stimuli and new-instance transfer stimuli. Illustrated here is the data from a sample run which has an accuracy of 100% in test trials. Tetrahedrons are for illustration purposes only.

5.3.2.3 Quantifying abstraction in hidden layer representation

We quantify abstraction in hidden layer representation by computing a parallelism score (PS) for each abstract variable as in previous literature (Bernardi et al., 2020; Ito et al., 2022). This is essentially the cosine similarity between coding directions of a given abstract variable. A PS equal to 1 means that the coding directions are perfectly parallel. Conversely, a PS of 0 indicates orthogonal coding directions. For a stimulus defined by shape, colour and texture (s,c,t), we can extract hidden layer activation as a vector $f_{s,c,t} \in R^{20}$. For any given shape dichotomy $(s¿¿1,s_2)¿$, we compute the coding direction v conditional on a colour-texture combination vector (c,t) as

$$v\left(s_1,s_2\middle|(c,t)\right)=f_{s_1,c,t}-f_{s_2,c,t}$$

Then we compute the cosine similarity between two coding directions.

$$\cos\left(v_1,v_2\right)=\frac{v_1\cdot v_2}{\|v_1\|\times\|v_2\|}$$

where $\cdot$ indicates dot product. If coding directions are parallel, the value would be 1.

For any given pair of two shapes, there are 4*4 colour-texture combinations, yielding 16 coding directions. We compute the absolute cosine similarity between any two out of all 16 coding directions. The parallelism score of shape dichotomy (s_1,s_2) is defined as the average of these absolute cosine similarities using

$$PS\left(s_1,s_2\right)=\frac{1}{\binom{4*4}{2}}\times\sum_{(c,t)_1\neq(c,t)_2}\cos\left[v\left(s_1,s_2\middle|(c,t)_1\right),v\left(s_1,s_2\middle|(c,t)_2\right)\right]$$

Then the PS for shape is given by averaging the PS over all possible shape dichotomies

$$PS_{shape}=\frac{1}{\binom{4}{2}}\times\sum_{s_1\neq s_2}PS(s_1,s_2)$$

$$¿\frac{1}{\binom{4}{2}}\times\sum_{s_1\neq s_2}\frac{1}{\binom{4\times4}{2}}\times\sum_{(c,t)_1\neq(c,t)_2}\cos\left[v\left(s_1,s_2\middle|(c,t)_1\right),v\left(s_1,s_2\middle|(c,t)_2\right)\right]$$

Similarly, the PS for colour pair is defined as the average PS over all possible colour dichotomies. Note that there are two correlated colours within each colour pair, e.g. $CP_1=(¿,¿),CP_2=(¿,¿)$. Hence, the colour dichotomies are all possible combinations of CP_1 and CP_2.

$$PS_{colour\ pair}=\frac{1}{2\times2}\times\sum_{c_1\in CP_1,c_2\in CP_2}\frac{1}{\binom{4*4}{2}}\times\sum_{(s,t)_1\neq(s,t)_2}\cos\left[v\left(c_1,c_2\middle|(s,t)_1\right),v\left(c_1,c_2\middle|(s,t)_2\right)\right]$$

5.3.2.4 Observed geometry and abstraction in hidden layer representation

To quantitatively test our prediction, we run two analyses. In the first analysis, we compute the RDM of both models (Figure 5-6A&C right panels) and examine how well they correlate with the observed hidden layer geometry. In the second analysis, we compute PS for shape and colour pair at the time of initialization and the final epoch.

Consistent with our prediction, if the network undergoes the full training regime, its hidden layer representation has a higher correlation with the parallel tetrahedron model than the seven-simplex model (Figure 5-7A), and the PS scores for both abstract variables are higher in the final epoch compared to initialization if the network receives full training (Figure 5-7B). Lack of pre-training on the relational structure lowers the correlation between hidden layer representation and the parallel tetrahedron model (Figure 5-7A), as

well as the PS scores (Figure 5-7B). In accordance with results on accuracy, training curricula do not affect hidden layer geometry.

More importantly, hidden layer geometry is relevant for generalization. We focus on the test accuracy in the new-instances transfer setting at the final epoch in the full training regime. A higher correlation between hidden layer geometry and parallel tetrahedron model scales with higher accuracy during test trials (Figure 5-7C). The same trend is found in PS for colour pair, but interestingly, not shape (Figure 5-7D). The PS for shape is close to one on average. In marked contrast, the PS for colour is lower and spans a broader range. A direct consequence is that the network does not make any shape errors in the final epoch, all errors are colour errors. This also explains why the test performance is only correlated with PS for colour pair.

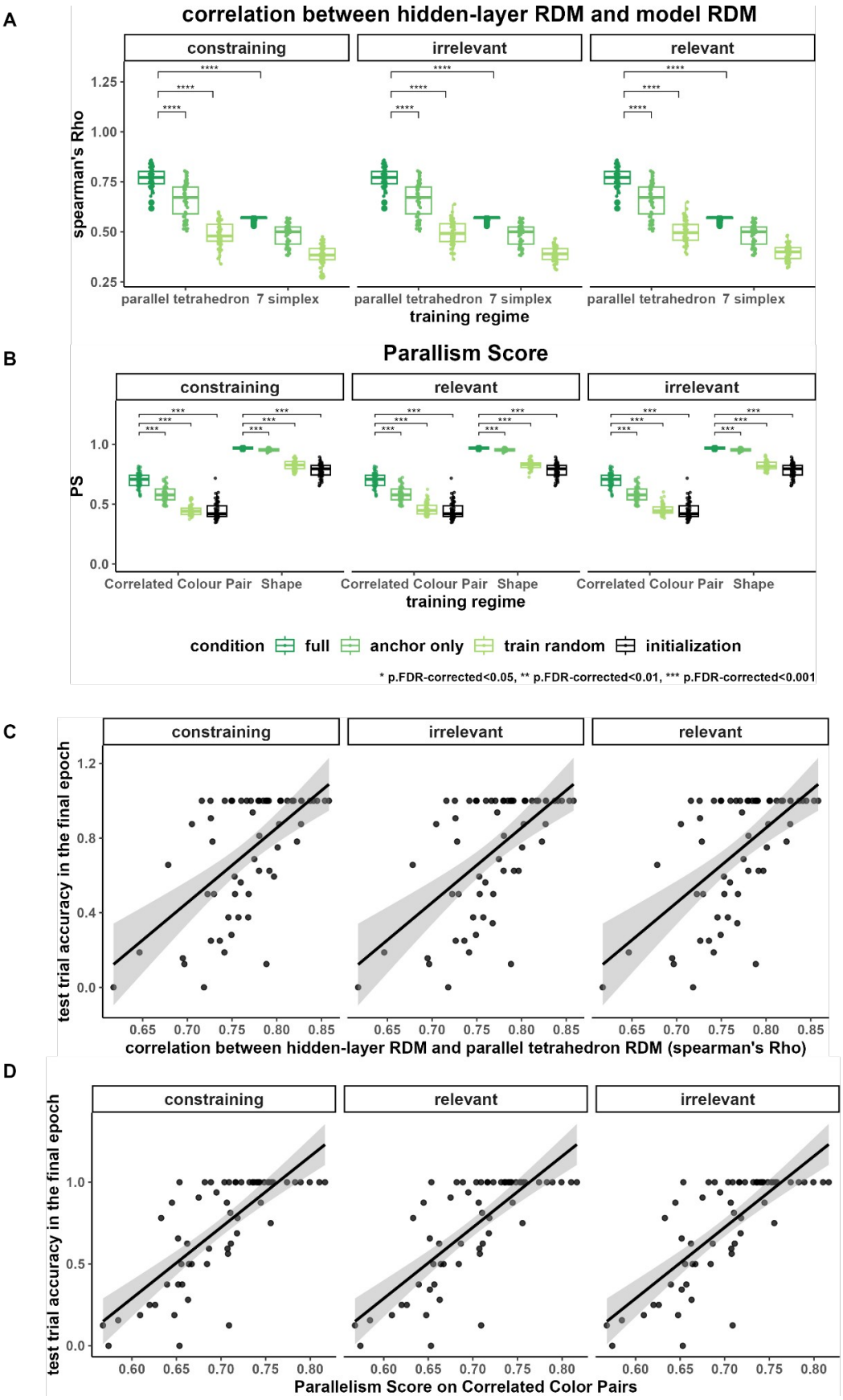

A
correlation between hidden-layer RDM and model RDM
constraining
irrelevant
relevant
spearman's Rho
1.25
1.00
0.75
0.50
0.25
parallel tetrahedron 7 simplex
training regime
B
Parallism Score
PS
1.0
0.5
0.0
Correlated Colour Pair Shape
condition full anchor only train random initialization
* p.FDR-corrected<0.05, ** p.FDR-corrected<0.01, *** p.FDR-corrected<0.001
C
test trial accuracy in the final epoch
1.2
0.8
0.4
0.0
0.65 0.70 0.75 0.80 0.85
correlation between hidden-layer RDM and parallel tetrahedron RDM (spearman's Rho)
D
0.60 0.65 0.70 0.75 0.80
Parallelism Score on Correlated Color Pairs

Figure 5-32. Observed hidden layer representations in different runs and their correlation with transfer performance in test trials in the final epoch. (A) Correlation between observed hidden layer RDM and model RDM. Asterisks indicate the results of two-sample (for between-regime comparisons) or pair-wise (for between-model RDM comparisons) comparisons using t-tests on Fisher z-transformed correlation coefficients. (B) PS score for colour pair and shape. Asterisks indicate the results of two-sample (for between-regime comparisons) or pair-wise (for between-model RDM comparisons) comparisons using Wilcoxon signed-rank test. (C-D) Correlation between geometry and test trial accuracy in the final epoch in the full training regime. Geometry is measured by (C) correlation with parallel tetrahedron model or (D) PS score on correlated colour pair.

5.4. Summary and Discussion

The current project aims to understand how humans learn about the covariance structure of feature variables and what supports generalization of learned covariance structure to either new instances or new response space. The first question is probed by comparing the benefit of different training curricula whilst the second is probed by deriving a theoretical prediction of the representation format subserving structure transfer using neural network models.

In experiment 1, we find that participants can learn and generalize the covariance structure among feature variables to a new response space. Blocking features that define the covariance structure significantly benefit learning and transfer performance in participants. Furthermore, temporal autocorrelation provides a context for learning, as illustrated by different contextual errors participants make under different curricula. We also train an artificial agent to perform the task and find that it cannot make counterfactual inferences to solve the

task as humans do and that temporal autocorrelation manipulation does not affect its behaviour.

In experiment 2, we design a new task variant of experiment 1 and eliminate the counterfactual inference component. The network could in theory solve the transfer problem. However, across multiple runs, test performance shows great variability albeit ceiling performance in training and anchor trials. We derive a theoretical model of representation that support structure transfer and test our prediction with the simulated data. Theoretically, we argue that an abstract representation in which context and structure is factorised is required for generalization. Empirically, this is backed up by correlations between measures of observed hidden layer geometry and test accuracy in the final epoch. Relatedly, this also echoes the hippocampal results reported in Chapter 3: the two rules—each mapping a feature dimension to a spatial axis—are represented in two parallel high-dimensional manifolds, indicating a factorisation between the rule identity and the magnitude axis (which captures progression along the spatial axis).

However, with a standard back-propagation learning algorithm, the naïve MLP does not show sensitivity to the temporal organization of training examples. Hebbian learning has been proposed as a mechanism for extracting structure from temporal statistics. Future work still needs to be done to test if MLP embedded with a Hebbian learning rule can replicate the findings in human learners.

Chapter 6 General Discussion

This thesis focused on how human learners form abstract representations throughout learning and later used these representations in a compositional fashion that enables flexible generalization behaviour. In the introduction, we highlight factorization and compression as two core properties of abstract representations. Factorization refers to the separation of underlying data-generating factors within the representational space. For example, average temperature and rainfall can be treated as distinct factors that together shape the distribution of tree types across different regions. Similarly, factorized representation of common structure and problem contexts also allow for multi-task learning and analogical transfer. Compression, on the other hand, involves discarding irrelevant features from the representation. For instance, the spatial position of objects is typically irrelevant for object classification and is thus omitted from the abstract representation for object recognition. Together, they ensure compositionality, representations can be either flexible combined to make novel inferences (i.e., binding individual knowledge pieces) or flexibly applied to novel contexts (i.e., binding structure to specific problems).

Chapter 2 and 3 leveraged an established task (Dekker et al., 2022) and provided evidence of abstract neural representations that are used in a compositional fashion to solve novel inferences. The task involved learning of two distinct rules, each mapping a set of discrete features (colours/shapes) onto a one-dimensional continuous axis (x-axis/y-axis). Importantly, rules were taught simultaneously.

Participants were provided the correct treasure locations of training stimuli without explicit instructions on how each feature was predictive of spatial locations. Therefore, learners had to factorise the rules from learning experiences.

Chapter 2 showed that different brain regions employed different representation strategies for abstraction. First, to allow for efficient readout, the rules could be compressed to two one-dimensional axes in the representation space. While this was true for posterior parietal cortex (PPC) and early visual cortex (V1), we found that this was not the case for hippocampus and ventral medial prefrontal cortex (vmPFC). Representation in these regions maintained high-dimensional, enabling feature separation. Using vector addition analyses, we confirmed that representations of test stimuli can be predicted from the weighted linear summation of training stimuli representations in vmPFC, PPC, and V1. Moreover, in line with the dimensionality results, vector addition in vmPFC followed a high-dimensional composition model in that the relevant training stimuli contributed to the prediction more than the irrelevant ones. For example, prediction of yellow whale relied more on the representations of yellow elephant and purple whale than other training stimuli. In comparison, in low-dimensional regions such as PPC and V1, the weights of the irrelevant training stimuli were further modulated by the distance to the relevant ones on the ground-truth map.

To explore the cause as well as the advantage of such high-dimensional representations, we ran simulations with feedforward neural networks that had one hidden layer (Chapter 3). We demonstrated two cases where such high-dimensional representation could emerge. The first one was the need to maintain feature separation, which was implemented by training the models with an

additional loss to reconstruct the input (i.e., having an auto-encoder head). The second one was having high-dimensional representations prior to task learning, which was elicited through unsupervised pretraining on task stimuli. We speculated that the latter was the cause behind the high-dimensional representations observed in hippocampus and vmPFC in our data because the task stimuli were composed from common animals and colours. In other words, participants may come to our task with prior representations in which these commonly seen colours and shapes were well separated. We further showed that such high-dimensional representation allowed for fast, flexible learning of new readout rules. When the features were represented in a high-dimensional format in the hidden layer, learning of new rules that shuffled the mapping from feature to spatial locations can be realised by merely adapting the readout weights connecting hidden layer and the output layer. Note that we did not employ any weight-freezing procedure, this dynamic arises naturally during learning. In comparison, if initial learning led to a compressed representation of features to two low-dimensional spatial axes, learning of shuffled mapping would require changing the encoding weights as well as the readout weights.

Apart from the high dimensionality, hippocampus also demonstrated a higher-level of abstraction of task rules that was not directly relevant to the task solution. As the two rules corresponded with two orthogonal spatial axes, a natural solution would be to maintain as such in the representation space. However, hippocampus represented the two rules as two parallel manifolds as if the two rules were registered to a common magnitude axis that described the progression along the axes. Such representations can be viewed as the factorised representations of a common relational structure (i.e., the magnitude or progression along axis) and contexts (i.e., the

individual rules). While there were individual variations in whether the rules were represented parallel to each other, a large portion of the participants do so. There was no clear benefit of parallel representation in this task, and indeed, the geometry was not related to the task accuracy. A future direction would be to identify the factor contributing to the individual variations and design computational models that can replicate such individual variations. Intuitively, we think this may be related to the role of hippocampus in representing common structure across experiences. While hippocampus was not involved in the vector addition operation for binding the two rules, it may be involved in binding the relational structure (i.e., magnitude) with the individual rules via operations like tensor product. Empirically, we plan to further test the replicability of such higher-level abstraction in the cross-context generalization task shown in Chapter 4. In this cross-context task, participants were trained on two training maps thereby creating four rules that can be mapped to a common relational structure. Our pilot studies showed that the learning and generalization performance was not impaired by the introduction of multiple maps, making it a promising testbed for further replicating our current results.

In Chapter 5, we further explored the question of the factorised representation of structure and its transfer to new situations. We modified the treasure hunt task to a name prediction task where the stimuli features indicated mapping to discrete labels. Unbeknownst to participants, there were hidden relational rules in the task: 1) different animals would have different labels; 2) there were colour pairs that would have the same label (e.g. blue whale and yellow whale would both be associated with the label ‘akar’, blue duck and yellow duck would both be associated with the label ‘hux’). After training stimuli-label associations, participants successfully transfer

the underlying structure to aid prediction on a new set of labels. We ran a simulation study that introduced two label-sharing colour pairs (i.e., blue and yellow whale shared one label, green and purple whale shared another label) and asked if relationship between colour pairs learned in one context (i.e., animal) can be generalized to another (i.e., new animals). Unlike the treasure hunt task, where abstraction of the 'magnitude' structure was not essential, in this case, abstraction of the 'sameness' structure was crucial for generalization. We measured the degree of abstraction by computing the cosine similarity between the representations of the colour pairs in different shapes. The rationale was that a factorised representation of the "sameness" structure, would lead to orthogonal representations of context variables and structure variables, making the instantiation of the same structure in different context parallel to each other, allowing for information sharing. Indeed, we found that the degree of abstraction observed hidden layer geometry was correlated with transfer performance accuracy.

Additionally, we investigated the benefits of temporal regularities in learning structures in Chapter 5. In the task, we created stimuli from the factorial combinations of three feature dimensions (colour, shape, texture). As introduced above, both colour and shape were relevant for the rules mapping the stimuli to labels, therefore they were the structure-defining features. We showed that curricula that allowed these structure-defining features to vary slowly promoted structure learning and transfer. This is in line with the known reliance on temporal statistic to discover latent structure in the data reviewed in the introduction. The naïve standard feedforward neural network we implemented was not sensitive to this temporal regularity in the training curriculum. One future direction of work would be to explore if we use Hebbian plasticity mechanisms commonly used in temporal

learning to model the human participants' sensitivity to training curricula.

The goal of this thesis is to contribute to the understanding of representations underlying compositional generalization. Together, our results provide further empirical support for the importance of abstract representation in generalization behaviour. Our work showed that factorised representations of data-generating factors, such as independent task rules or common structure and context, are crucial for generalization. We showed evidence of a vector addition process supporting the binding of independent task rules to solve novel inferences compositionally in vmPFC, PPC and V1. While hippocampus is not directly involved in such a process, it formed abstraction of the shared relational structure between rules, indicating a division of labour between regions. The variations in abstraction in different cortical and subcortical areas are likely linked to their representational demands and functional roles in the task. Through computational simulations, we show that variations in representational dimensionality contribute to different learning dynamics when a new task is introduced. This suggests that a useful, abstract representation needs to strike a balance between short-term efficiency and long-term flexibility. Several pieces of work presented in this thesis represent on-going effort to further explore these questions, which hopefully not only informs how the biological brain efficiently adapts to the changing environment, but also guides external efforts, such as curriculum design, facilitate such adaptive behaviour.

Appendix A: Supplementary Materials for Chapter 2

A.1 Hypothesis about rule representation

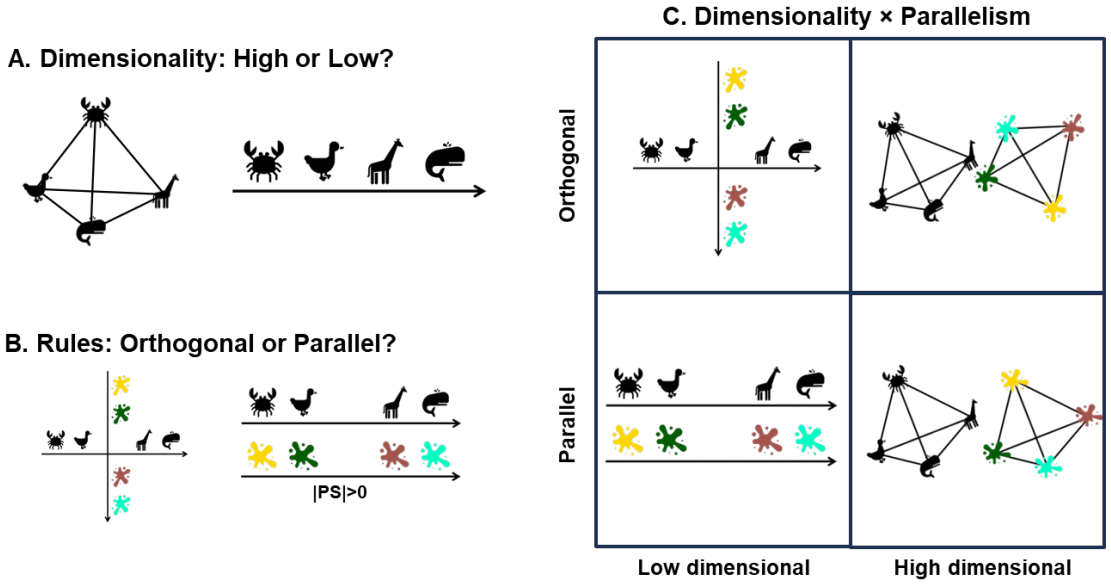


Figure S2-1. Potential strategies for rule representations. (A) Dimensionality: Before learning, the representation of discrete features was high-dimensional, allowing for feature separation (left). After training, the representation could remain high dimensional, or could compress to a low-dimensional axis (right) that reflect the correspondence between features and axis locations. (B) Relationship between rules: Space could be used as a scaffold to represent the rules. If so, the rules would be represented orthogonally (left). Alternatively, a shared relational primitive, magnitude, could be used to describe the progression along the linear axis (left-ness and top-ness). If so, the rules would be represented in parallel (right, example of top of y aligned to left of x). (C) Space of potential strategies derived from the two characteristics of rule representation in (A) and (B).

A.2 Participant Exclusion based on behaviour consistency

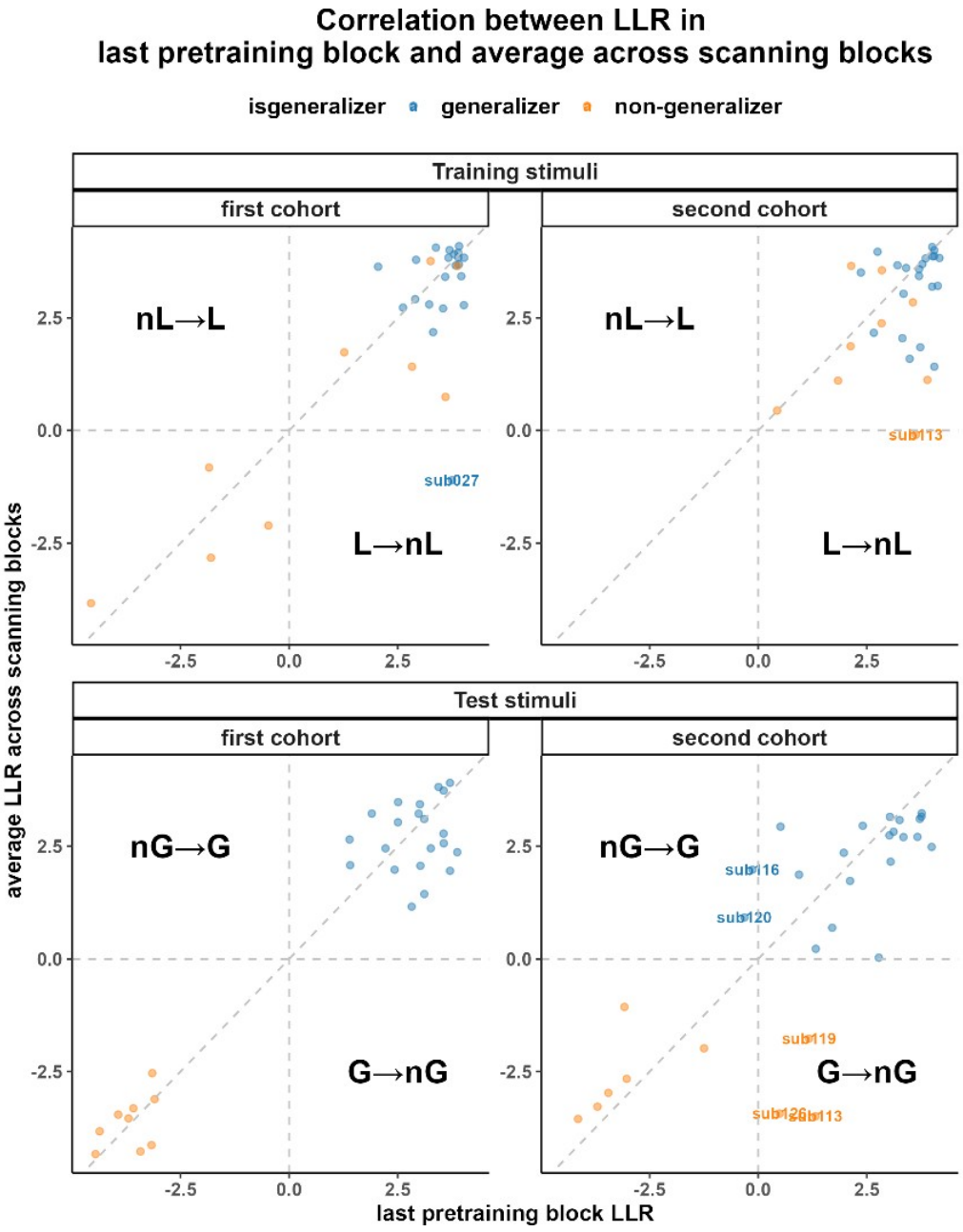


Figure S2-2. Participant Exclusion based on inconsistent (deteriorating) behaviour in the scanner compared to pretraining. The plot showed the relationship between the LLR from the last pretraining block and the LLR in the scanner for training stimuli (top rows) and test stimuli (bottom rows) as well as for different cohorts separately. Each dot represented one participant.

A.3 Effect of multivariate noise normalisation

To examine the effect of multivariate noise normalisation in a way that is independent from the analyses driven by our hypotheses, we computed two metrics as proxies of signal-to-noise ratio in multivariate data: discriminability and stability.

Discriminability concerns the degree to which the activity patterns of different stimuli can be reliably discriminated when taken from different runs. This is equivalent to asking if there is a greater similarity between activity patterns of the same stimulus from different tuns than between activity patterns of different stimuli from different runs. To this end, we averaged the activity patterns of each stimulus from odd runs and even runs, yielding two activity pattern matrices $X_{odd}=[x_{1,odd}, x_{2,odd},\ldots, x_{25,odd}]^T$ and $X_{even}=[x_{1,even}, x_{2,even},\ldots, x_{25,even}]^T$, where $x_{j,odd}$ and $x_{j,even}$ are the activity patterns of stimuli j averaged across odd and even runs respectively. We then computed a cross-run neural RDM, and the entry from the i-th row and the j-th column, $RDM_{i,j}$, is computed as:

$$RDM_{i,j}=1-pearsonr\left(x_{j,odd}, x_{k,odd}\right)$$

We then computed the spearman's rank correlation between this cross-run neural RDM and a model RDM in which all diagonal entries are 0 and all non-diagonal entries are 1. This correlation coefficient is used as a metric for discriminability.

Stability asked if the representation structure, measured by neural RDM, remained stable across runs. We computed neural RDMs using the average neural activity matrices: X_{odd} and X_{even}. Then, we computed the spearman's rank correlation between the lower

triangular parts of the odd-run RDM and even-run RDM as the metric for stability.

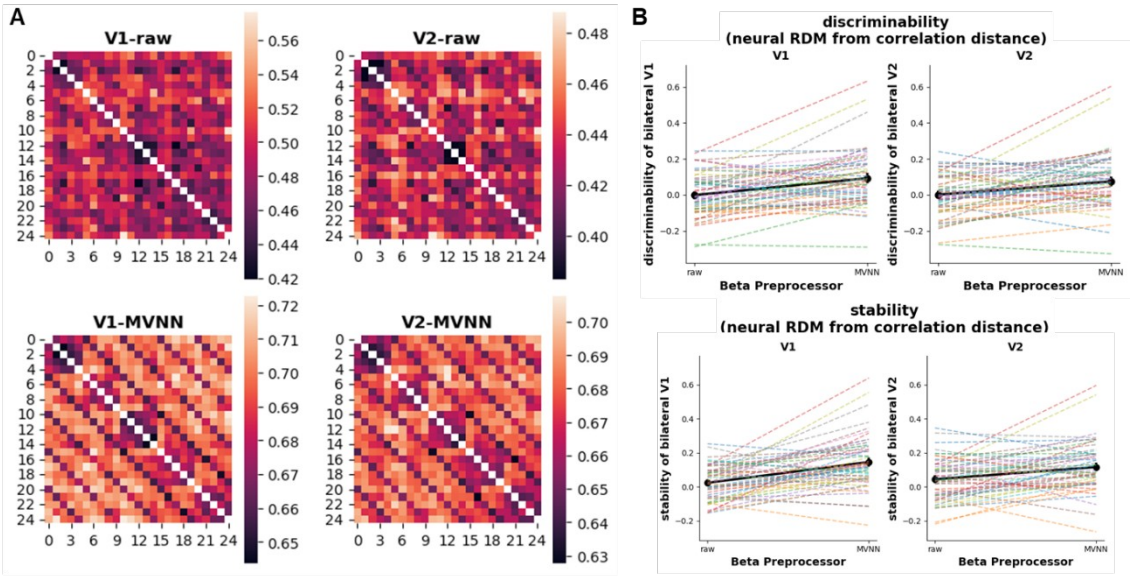


Figure S2-3. Effect of multivariate noise normalisation (MVNN) in perceptual regions (V1&V2). (A) neural RDM for all the stimuli averaged across generalizers from V1 (left columns) and V2 (right columns) before (top rows) and after (bottom row) MVNN. The neural RDM was ordered by stimuli colour and then stimuli shape. That is, indexed 0-4 were yellow animals (crab, duck, elephant, giraffe, whale), indexed 5-9 were green animals (crab, duck, elephant, giraffe, whale) etc. (B) Quantifying improvement with discriminability (top) and stability (bottom) metrics. Coloured dashed lines represent individual generalizers, black solid line represent mean.

A.4 Retrieval pattern analysis sanity checks

We performed two sanity checks as described in *2.3.11 Retrieval pattern analysis: Predicting test stimulus representation from training stimuli*. The first one was the prediction quality in held-out measured by coefficient of determination (R^2) which asked if the retrieval patterns fitted in one half of data (e.g. odd runs) can be applied to the held-out half of data (e.g. even runs). The second one asked if the weights of the relevant training stimuli (β_{rel}) were above zero and higher than the weights of the irrelevant training stimuli (β_{irr}).

We can see that only V1, superior parietal gyrus (SPG), and hippocampus (HPC) had above-chance prediction quality, indicating potential involvement in vector addition composition. For V1, both generalizers and non-generalizers passed the sanity checks, indicating that V1 representation followed the vector addition hypothesis of composition. This is likely due to V1's role in processing the visual features (i.e., colours and shapes) of the stimuli. However, in SPG and vmPFC, only the generalizers satisfied the constraints that the weights for relevant training stimuli were higher than those of irrelevant training stimuli. Therefore, we arrived at the conclusion that V1 in both generalizers and non-generalizers showed vector addition composition, and that only the generalizers' SPG and vmPFC showed vector addition composition.

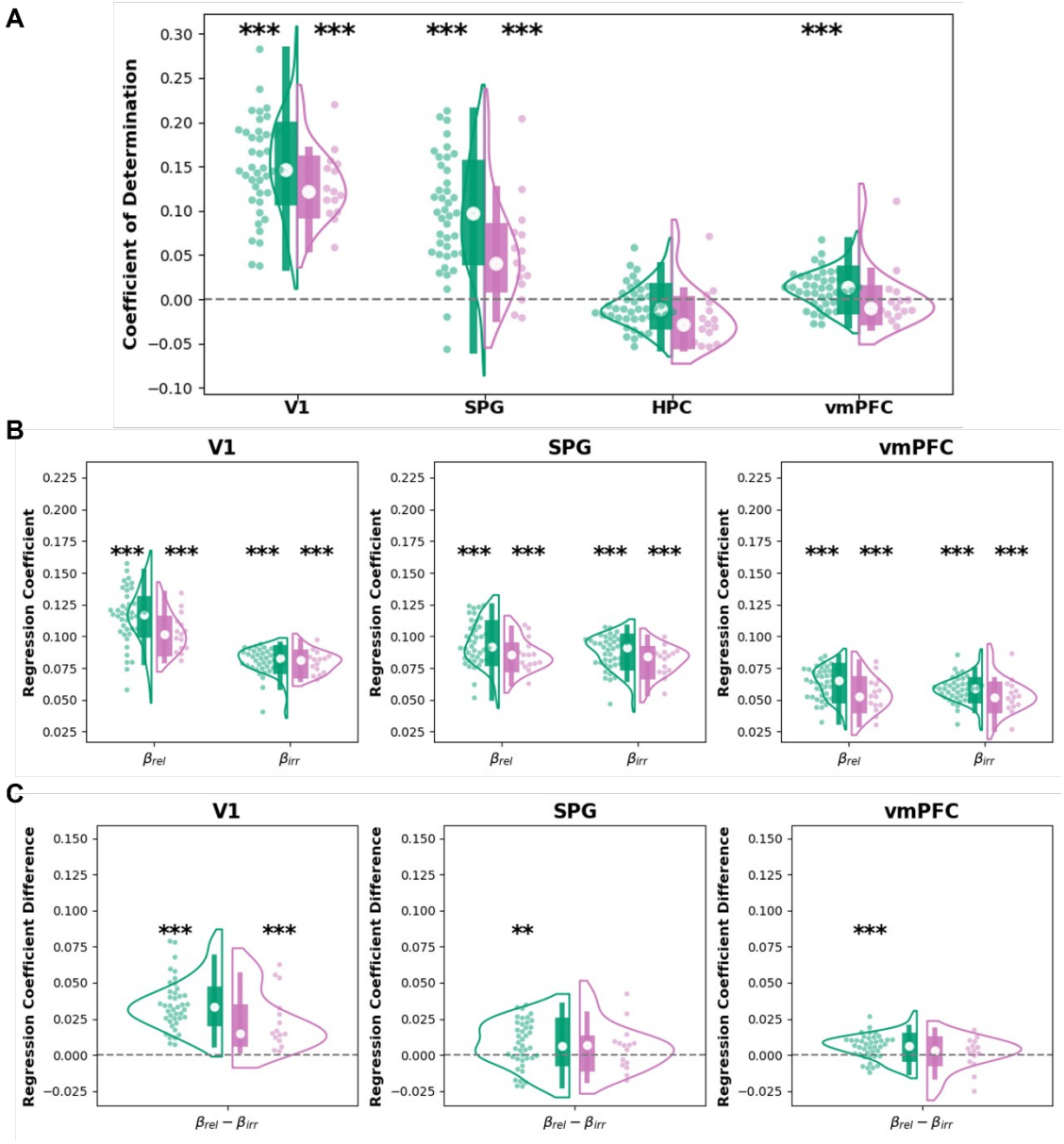


Figure S2-4. Sanity checks of retrieval pattern analysis. (A) Prediction quality of the retrieval patterns in held-out data measured by coefficient of determination R^2 in all four ROIs. An R^2 of zero indicated that prediction was at chance, and R^2 below zero indicated that the prediction worse than chance. Only R^2 that was significantly above zero indicated vector addition composition. (B) Average weights of relevant (β_{rel}) and irrelvant training stimuli (β_{irr}) in the retrieval patterns for the three ROIs that showed vector addition composition. (C) Difference between weights of

relevant and irrelvant training stimuli. *Notes*: Asterisks indicated results of permutation tests, *p<0.05, **p<0.01, ***p<0.001. In the violin plots (D-F), dots represented individual participants. Box indicated the quantiles in each participant group, while whiskers extended to points within 1.5 IQRs of the lower and upper quartile.

A.5 Stimuli group decoding

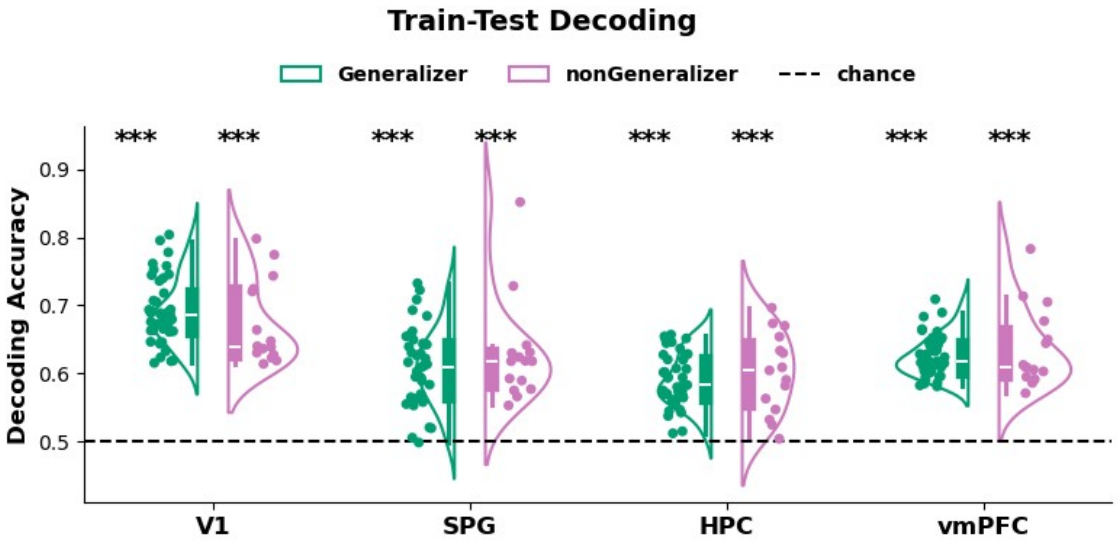
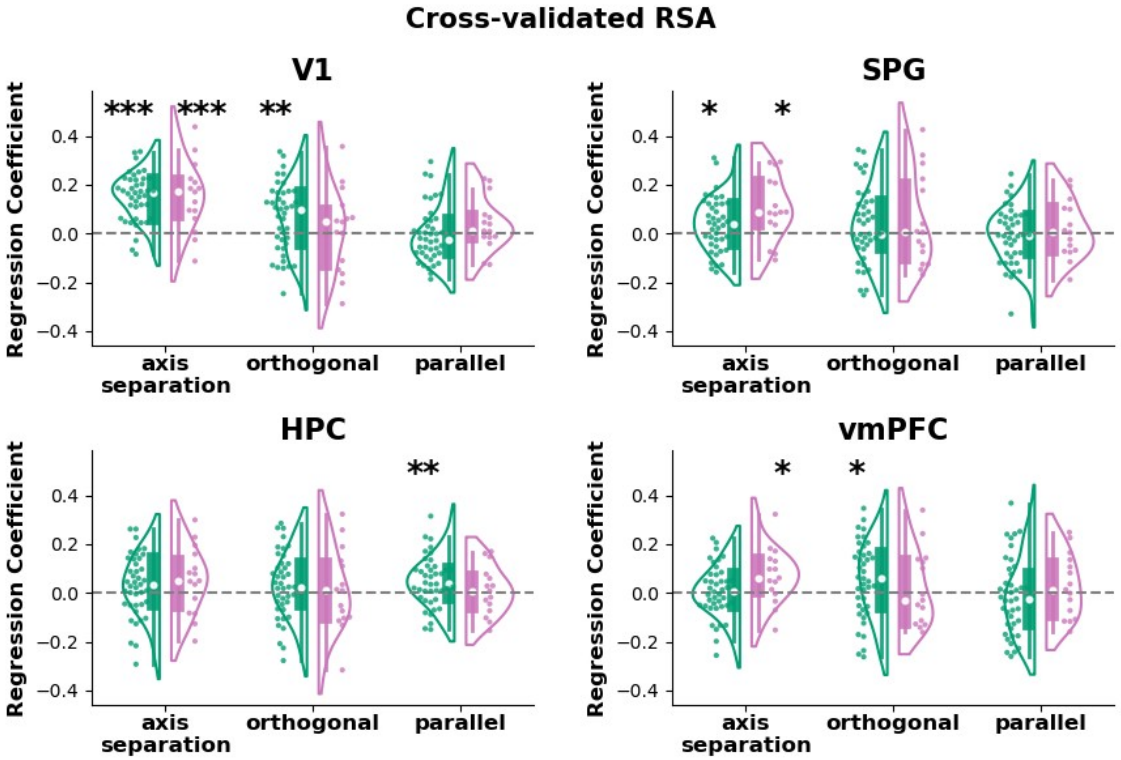


Figure S2-5 Stimuli group decoding. Decoding stimuli group identify (training vs test) from stimuli representation. Dots represented the decoder's mean evaluation accuracy across 20 random seed for one participant. Box indicated the quantiles in each participant group, while whiskers extended to points within 1.5 IQRs of the lower and upper quartile. Asterisks indicated results of permutation tests, *p<0.05, **p<0.01, ***p<0.001.



A.6 Testing parallel representation with cross-validated RSA

Figure S2-6. Testing parallel representation with cross-validated RSA. Cross-validated RSA comparing neural RDM to axis separation, orthogonal and parallel model RDMs in all ROIs. Dots represented individual participants. Box indicated the quantiles in each participant group, while whiskers extended to points within 1.5 IQRs of the lower and upper quartile. Asterisks indicated results of permutation tests, *p<0.05, **p<0.01, ***p<0.001.

A.7 Whole-brain searchlight RSA results

Table S2-1. Searchlight RSA results: regressing all stimuli neural RDM on three model RDMs: feature RDM, location RDM, train-test separation RDM.

Notes:
[1]The reported second-level results for this regression searchlight employed a more stringent cluster-defining threshold at p<0.0001 otherwise the cluster would be too large for some of the results.
[2] Clusters at which the peak passed whole-brain FEW correction were also reported.
[3]Anatomical labels were from AAL. When AAL labels were not available, the Harvard-Oxford cortical and subcortical atlas was used.

A. feature RDM

Anatomical Label	Peak coordinates (MNI)			Peak tvalue	Peak p_{FWE}	Cluster	
	x	y	z			Size	p_{FWE}
Lingual_R	15	-86	-10	16.480	<0.001	9425	<0.001
Calcarine_R	13	-96	-2	16.382	<0.001		
Occipital_Inf_R	30	-86	-2	16.340	<0.001		
Temporal_Mid_L	-58	-51	11	5.875	0.006	236	<0.001
Precuneus_L	-3	-56	61	5.070	0.049	120	0.009
Frontal_Inf_Oper_L	-53	17	18	5.018	0.056	221	0.001
Frontal_Inf_Tri_L	-40	17	26	4.995	0.059		
Precentral_L	-20	-9	43	4.632	0.144	52	0.051
Precentral_L	-25	-4	48	4.525	0.185		

B. location RDM

Anatomical Label	Peak coordinates (MNI)			Peak t value	Peak p_{FWE}	Cluster	
	x	y	z			Size	p_{FWE}
Calcarine_L	-5.0	-86.0	10.5	10.193	<0.001	1948	<0.001

Calcarine_R	12.5	-78.5	8.0	9.465	<0.001	1948	
Precentral_L	-40.0	-18.5	58.0	5.483	0.012	96	0.023

C. train-test separation RDM

Anatomical Label	Peak coordinates (MNI)			Peak t value	Peak p_{FWE}	Cluster	
	x	y	z			Size	p_{FWE}
Parietal_Inf_L	-35	-51	40.5	7.847	0.000	4848	0.000
Precuneus_L	-7.5	-58.5	45.5	7.661	0.000		
Parietal_Inf_L	-25	-58.5	40.5	6.950	0.000		
Frontal_Mid_2_L	-37.5	46.5	-9.5	7.313	0.000	5129	0.000
Frontal_Mid_2_L	-37.5	34	33	7.311	0.000		
Frontal_Inf_Tri_L	-42.5	34	18	7.187	0.000		
Supp_Motor_Area_R	12.5	-13.5	58	6.095	0.004	175	0.001
Precentral_R	27.5	-6	53	5.433	0.022		
Frontal_Sup_2_R	22.5	-3.5	63	4.491	0.232		
Frontal_Mid_2_R	35	51.5	15.5	5.880	0.007	225	0.000
Frontal_Mid_2_R	42.5	44	13	5.058	0.059		
Hippocampus_R	37.5	-11	-14.5	5.726	0.010	37	0.075
Brain stem	2.5	-31	-29.	5.269	0.034	28	0.107

			5				
Paracentral_Lobule_L	-15	-26	73	4.912	0.086	50	0.047
Cingulate_Mid_R	7.5	-26	48	4.773	0.121	136	0.004
Paracentral_Lobule_L	-7.5	-31	53	4.515	0.220		
Cingulate_Mid_L	-7.5	-33.5	43	4.345	0.316		
white matter	20	-21	33	4.768	0.122	108	0.008
white matter	27.5	-23.5	20.5	4.472	0.242		
Insula_R	37.5	-11	18	4.451	0.253		
Insula_L	-42.5	-1	0.5	4.764	0.124	35	0.081
Central Opercular Cortex	-45	1.5	-7	4.165	0.445		
Frontal_Sup_Medial_R	15	49	0.5	4.719	0.138	57	0.037
ACC_pre_R	15	39	0.5	4.536	0.210		

Table S2-2. Searchlight RSA results: regressing test stimuli neural RDM on two model RDMs: feature RDM and location RDM

Notes:
[1]The reported second-level results for this regression searchlight employed a cluster-defining threshold at p<0.001.
[2] We also reported some subthreshold clusters in grey if their contralateral cluster passed cluster correction or voxel-wise correction.
[3]Anatomical labels were from AAL. When AAL labels were not available, the Harvard-Oxford cortical and subcortical atlas was used.

A. feature RDM

Anatomical Label	Peak coordinates (MNI)			Peak t value	Peak p_{FWE}	Cluster	
	x	y	z			Size	p_{FWE}
Occipital_Mid_L	-22.5	-96	10.5	14.201	<0.001	5611	<0.001
Occipital_Inf_R	27.5	-91	-7	13.999	<0.001		
Calcarine_R	12.5	-96	-2	13.772	<0.001		

B. location RDM

Anatomical Label	Peak coordinates (MNI)			Peak t value	Peak p_{FWE}	Cluster	
	x	y	z			Size	p_{FWE}
Calcarine_R	13	-79	13	8.526	0.000	2945	0.000
Calcarine_L	-8	-86	11	8.114	0.000	2945	
Cuneus_R	15	-89	36	3.472	0.905	2945	
Precentral_L	-38	-24	63	6.071	0.003	454	0.004
Postcentral_L	-38	-36	63	4.740	0.111	454	
Parietal_Inf_L	-58	-24	48	4.047	0.479	454	
Temporal_Pole_Mid_L	-30	12	-35	5.068	0.049	32	0.656

Temporal_Pole_Mid_R	35	19	-35	4.896	0.075	39	0.598
Precuneus_R	13	-66	56	4.071	0.460	225	0.051
Parietal_Sup_R	28	-61	56	3.923	0.581	225	
Parietal_Sup_L	-23	-56	66	4.245	0.333	102	0.249

Table S2-3. Searchlight RSA results: correlation between pirate position neural RDM and location model RDM

Anatomical Label	Peak coordinates (MNI)			Peak t value	Peak p_{FWE}	Cluster	
	x	y	z			Size	p_{FWE}
Lingual_R	5.0	-81.0	-2.0	33.555	<0.001	21010	<0.001
Lingual_L	-7.5	-83.5	-2.0	29.789	<0.001	21010	
Lingual_R	20.0	-86.0	-7.0	26.481	<0.001	21010	
Location not in atlas	37.5	-6.0	-14.5	5.477	0.010	213	0.006
Fusiform_R	35.0	-6.0	-34.5	4.454	0.133	213	
ParaHippocampal_R	25.0	-1.0	-34.5	4.349	0.169	213	
OFCpost_L	-45.0	24.0	-14.5	4.977	0.036	76	0.043
Frontal_Mid_2_L	-42.5	24.0	43.0	4.857	0.049	52	0.067

Note:
[1]The reported second level result employed a more stringent cluster-defining threshold at p<0.0001.
[2]Anatomical labels were from AAL. When AAL labels were not available, the Harvard-Oxford cortical and subcortical atlas was used.

Appendix B: Supplementary Materials for Chapter 4

B.1 Procedure of pilot M2V2 and pilot M2V3

The procedures of pilot M2V2 and M2V3 were almost identical except that in pilot M2V3 the order of evaluation tasks was switched. In pilot M2V3, participants completed the treasure hunt task before the 2AFC task. As we are interested in using the 2AFC task for fMRI scanning, we hope that it would be the first time participants are asked to make inference on the unseen test stimuli during pretraining. Therefore, we made the design choice to switch the order of the 2AFC task and the treasure hunt task in pilot M2V3.

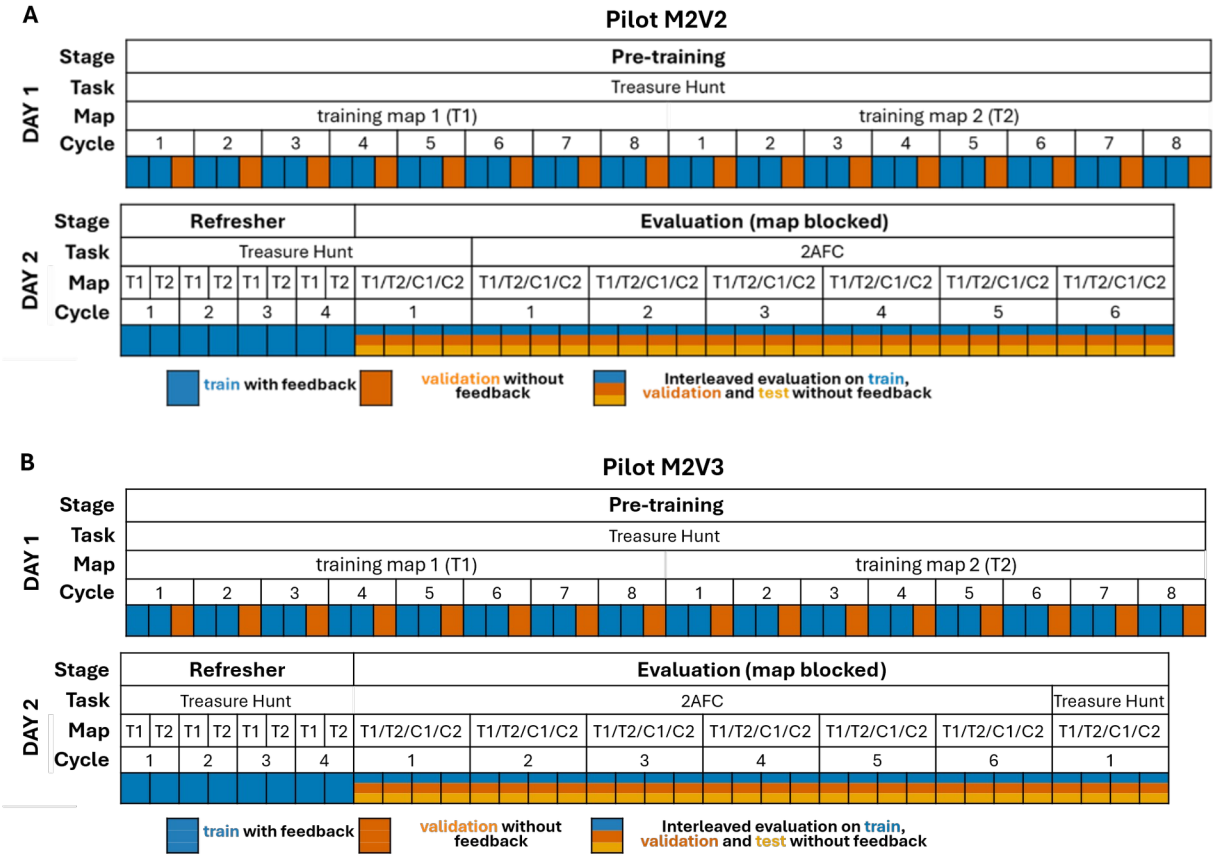


Figure S4-1. Procedure of multi-map pilot experiments. (A) pilot M2V2. (B) pilot M2V3.

B.2 Percentage of Missing trials in the 2AFC task

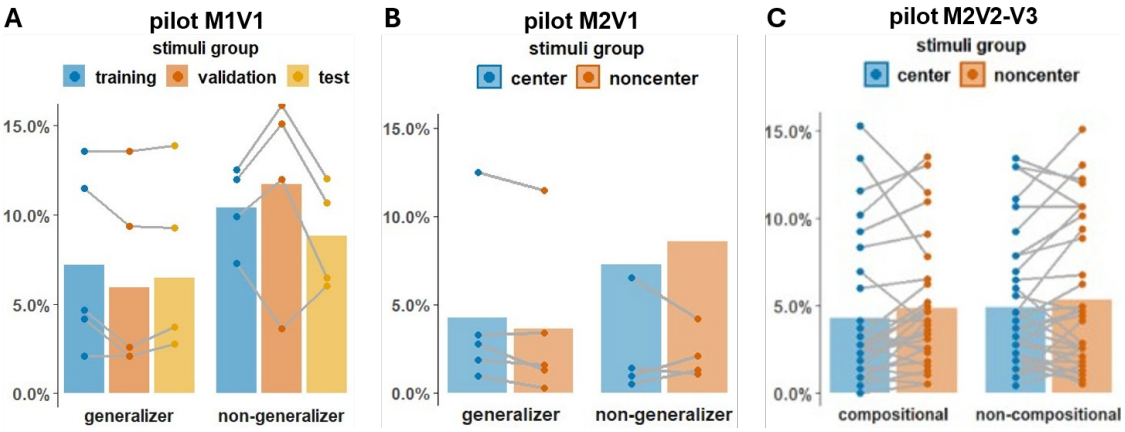


Figure S4-2. Percentage of Missing trials in different pilot experiments. Dots indicated individual participants.

Table S4-1. Percentage of missing trials in different pilot experiments. Note that there were very few participants in pilot M2V1, therefore, the standard error was quite large.

A. Pilot M2V1

Participant Group	Stimuli Group	Percentage (%) S.E.M.
generalizer	centre	4.259±2.098
	non-centre	3.594±2.030
non- generalizer	centre	7.222±5.026
	non-centre	8.542±6.417

B. Pilot M2V2-V3

Participant Group	Stimuli Group	Percentage (%) S.E.M.
compositional	centre	4.290±0.747
	non-centre	4.870±0.662
non-compositional	centre	4.930±0.652
	non-centre	5.346±0.733

B.3 Reclassifying stimuli as centre versus non-centre stimuli.

We examined participants' performance in the multi-map treasure hunt task as a function of stimulus type in the evaluation stage. This led us to discard the training versus non-training stimuli dichotomy and reclassify as centre versus non-centre stimuli.

Response error was measured by the distance between the participant's response and the correct location. First, we looked at five different stimulus types: training, validation and test stimuli from the training maps, and centre and non-centre stimuli from the cross maps. In the pilot M2V1, we observed that non-generalizers performed quite well on the centre stimuli in the cross maps. By design, the training stimuli were essentially the centre stimuli in the training maps. This led us to reclassify the stimuli as centre versus non-centre stimuli. The same pattern was observed in pilot M2V2-M2V3. As we have more data from the last pilots, we performed statistical tests to validate this observation.

We fitted a linear mixed effect model to trial-wise error with two independent variables (Participant Group, Stimuli Type), the interaction term, as well as random slope and random intercept for participants:

Distance ~ Participant Group + Stimuli Type + Participant Group × Stimuli Type + (1 + Stimuli Type | participant)

We set non-compositional generalizer as the reference level for the participant group variable. As for the stimuli type, the key contrast of our interest was the difference between centre in cross maps and training stimuli in training maps. Therefore, we set the training stimuli as the reference level.

Results (Table S4-2) revealed significant interactions between participant group and the non-centre stimuli vs training stimuli contrasts. But there was no significant interaction between participant group and the centre stimuli vs training stimuli contrast.

We further conducted paired t-tests between different stimulus types in different participant groups (results shown in Figure S4-3) as post-hoc tests. First, no significant difference was found among the three types of non-centre stimuli across maps: 1) validation stimuli from the training maps, 2) test stimuli from the training maps, 3) non-centre stimuli from the training maps. This was true in both compositional and non-compositional participants. Second, performance was significantly better in trials with centre stimuli than in trials with non-centre stimuli. This was true in both training maps (training stimuli error smaller than validation or test stimuli error) and cross maps (centre stimuli error smaller than non-centre stimuli error). This was also true for both compositional and non-compositional participants. Finally, non-compositional participants performed significantly better in trials with the centre stimuli from the cross maps than in trials with non-centre stimuli across maps. This indicated that compositional and non-compositional participants diverged more in the non-centre stimuli.

Together, this replicated and statistically confirmed the observation in pilot M2V1 that the compositional and the non-compositional participants diverged more in the non-centre stimuli across maps.

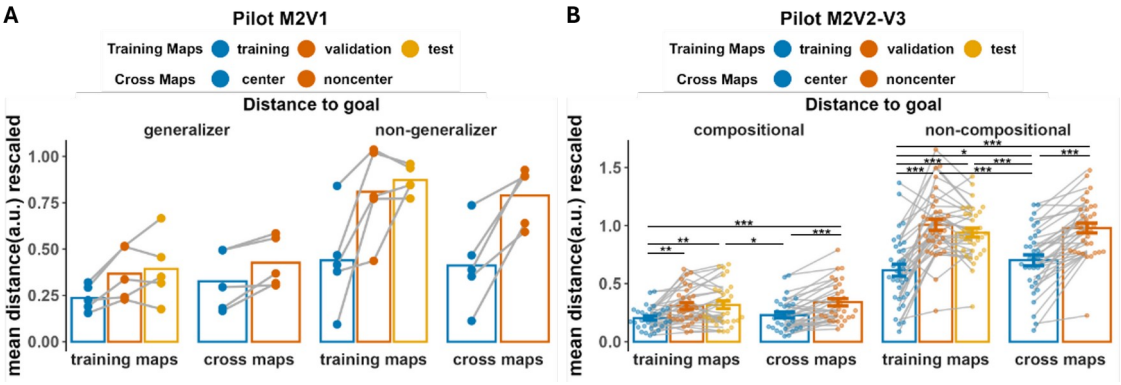


Figure S4-3. Performance of the treasure hunt task during the evaluation stage in (A) pilot M2V1 and (B) pilot M2V2-V3. Response error was measured by distance to the correct location. Bars represented the mean, and dots represented individual participants. Error bars in (B) represented the mean ± standard error. Asterisks in (B) represented results of paired t-test (Bonferroni-corrected): *p<0.05, **p<0.01, ***p<0.001.

Table S4-2 Effect of Five Stimuli Type × Participant Group on treasure hunt task evaluation performance (distance to correct location) in pilot M2V2-V3.

Regressors	Coefficients [C.I.]
(Intercept)	0.616 *** [0.537, 0.694]
Participant Group (G>NG)	-0.413 *** [-0.527, -0.298]
Stimuli Type (Val.>Train.)	0.391 *** [0.313, 0.469]
Stimuli Type (Test>Train.)	0.323 *** [0.252, 0.394]
Stimuli Type (CM-C>Train.)	0.086 ** [0.033, 0.139]
Stimuli Type (CM-NC>Train.)	0.363 *** [0.299, 0.427]
(G>NG) × (Val.>Train.)	-0.286 *** [-0.400, -0.172]
(G>NG) × (Test>Train.)	-0.209 *** [-0.313, -

	0.105]
(G>NG) × (CM-C>Train.)	-0.060 [-0.137, 0.017]
(G>NG) × (CM-NC>Train.)	-0.226 *** [-0.320, -0.132]
N (trial)	6400
N (participant)	64
AIC	7885.619
BIC	8061.484
R2 (fixed)	0.297
R2 (total)	0.420

*** p < 0.001; ** p < 0.01; * p < 0.05.
Abbreviations:
G: compositional generalizers; NG: non-compositional generalizers
Val.: validation; Train.: training
CM-C: cross-map centre; CM-NC: cross-map non-centre

The above results provided support for us to reclassify stimuli as centre versus non-centre stimuli rather than sticking to the previous training vs non-training stimuli dichotomy (Figure 4-10 in the main text). Therefore, we ran another linear mixed-effects model with three independent variables: Participant Group (compositional vs non-compositional: G vs nG), Stimuli Group (centre versus non-centre: C vs nC), and Map Type (cross maps vs training maps: CM vs TM). The interaction term, as well as random slope and random intercept for participants, were also included:

Distance ~ Participant Group + Map Type + Stimuli Group + Participant Group × Map Type + Participant Group × Stimuli Group + Map Type × Stimuli Group + Participant Group × Map Type × Stimuli Group + (1 + Stimuli Type | participant)

We set the non-compositional generalizers, centre stimuli, and training maps as the reference levels for the three independent variables respectively. Results were shown in Table S4-3.

Table S4-3 Effect of Participant Group × Stimuli Group × Map Type on treasure hunt task evaluation performance in pilot M2V2-V3

Regressors	Coefficients [C.I.]
(Intercept)	0.620 *** [0.548, 0.692]
Participant Group (G>NG)	-0.421 *** [-0.526, -0.315]
Map Type (CM>TM)	0.082 ** [0.033, 0.130]
Stimuli Group (C>nC)	0.371 *** [0.312, 0.430]
Participant Group × Map Type	-0.052 [-0.123, 0.020]
Participant Group × Stimuli Group	-0.248 *** [-0.334, -0.162]
Map Type × Stimuli Group	-0.093 ** [-0.154, -0.032]
Participant Group × Map Type × Stimuli Group	0.082 [-0.007, 0.171]
N (trial)	6400
N (participant)	64
AIC	7824.722
BIC	7926.183
R2 (fixed)	0.301
R2 (total)	0.421

*** p < 0.001; ** p < 0.01; * p < 0.05.

B.4 Evaluation stage performance in the multi-map 2AFC task as a function of centre versus non-centre stimuli

We fitted mixed effects models to trial-wise accuracy and response time (RT) with four independent variables (Participant Group, Difficulty, Stimuli Group, Trial), all the interaction terms, as well as random intercepts for participants:

Correct ~ Participant Group + Difficulty + Stimuli Group + Trial +

Participant Group × Difficulty + Participant Group × Stimuli Group + Difficulty × Stimuli Group + Participant Group × Trial + Difficulty × Trial + Stimuli Group × Trial + Participant Group × Difficulty × Stimuli Group + Participant Group × Difficulty × Trial

+ Participant Group × Stimuli Group × Trial + Difficulty × Stimuli Group × Trial + Participant Group × Difficulty × Stimuli Group × Trial + (1 | participant)

We set the non-compositional generalizer as the reference level for the participant group variable. And we set the non-centre stimuli as the reference level for the stimuli group variable. The continuous independent variables (Difficulty and Trial) were standardised before entering the regression. Trial-wise accuracy was fitted with a generalized linear mixed-effect model using a binomial link function. RT was fitted with a linear mixed-effect model. The group average results were plotted in Figure 4-10B in the main text, and the mixed effect model results were shown in Table S4-4.

Table S4-4 Effect of Participant Group × Stimuli Group × Difficulty on accuracy and RT in 2AFC task in pilot M2V2-V3.

	Accuracy	RT
(Intercept)	0.138	585.814 ***

	[-0.122, 0.399]	[555.225, 616.403]
Participant Group	1.220 ***	42.868
	[0.838, 1.602]	[-1.810, 87.546]
Difficulty	0.054 **	-1.172
	[0.020, 0.088]	[-3.584, 1.240]
Stimuli Group	0.332 ***	0.920
	[0.270, 0.394]	[-3.370, 5.209]
Trial	0.035	-3.319 *
	[-0.002, 0.072]	[-5.903, -0.735]
Participant Group × Difficulty	0.110 ***	-8.645 ***
	[0.053, 0.167]	[-12.157, -5.133]
Participant Group × Stimuli Group	0.003	-20.020 ***
	[-0.100, 0.106]	[-26.273, -13.767]
Difficulty × Stimuli Group	0.053	-4.228
	[-0.014, 0.120]	[-8.856, 0.400]
Participant Group × Trial	0.020	-0.968
	[-0.040, 0.080]	[-4.726, 2.789]
Difficulty × Trial	-0.012	-0.891
	[-0.046, 0.022]	[-3.289, 1.508]
Stimuli Group × Trial	-0.071 *	0.578
	[-0.132, -0.009]	[-3.701, 4.857]
Participant Group × Difficulty × Stimuli Group	-0.072	0.158
	[-0.185, 0.041]	[-6.581, 6.897]
Participant Group × Difficulty × Trial	0.027	1.747
	[-0.030, 0.084]	[-1.764, 5.257]
Participant Group × Stimuli Group × Trial	-0.003	2.653
	[-0.106, 0.100]	[-3.588, 8.894]
Difficulty × Stimuli Group × Trial	0.064	-2.484
	[-0.002, 0.131]	[-7.079, 2.111]
Participant Group × Difficulty × Stimuli Group × Trial	-0.135 *	2.988
	[-0.247, -0.023]	[-3.696, 9.672]
N (trial)	36501	36501
N (participant)	64	64
AIC	41614.102	467357.336
BIC	41758.689	467510.428
R2 (fixed)	0.096	0.014
R2 (total)	0.234	0.291

*** p < 0.001; ** p < 0.01; * p < 0.05.
95% Confidence intervals were shown in square brackets.

B.5 Behavioural consistency across the treasure hunt task and the 2AFC task

We tested whether participants' choices in the 2AFC were consistent with their responses in the treasure hunt task. More specifically, we were interested in the consistency of response in the non-centre stimuli trials. This would help determine if non-generalizers were responding randomly or if they were responding consistently according to their false belief of the treasure locations. To achieve this goal, we investigated whether participants' choices in the 2AFC task could be predicted from their responses in the treasure hunt task. First, we extracted their responses to each non-centre stimulus in the treasure hunt task during the evaluation stage. This gave us a response map of all the non-centre stimuli for each participant. Then, in each 2AFC trial, we computed the proximity of a given option to the location of the cue stimulus on this response map. Specifically, we computed the absolute distance on the x-axis and the y-axis separately:

$$prox\ y_{x,¿}¿ = \left| x_{¿} - x_{stimulus} \right|$$

$$prox\ y_{y,¿}¿ = \left| y_{¿} - y_{stimulus} \right|$$

Finally, we computed the difference between the proximity:

$$Dprox\ y_{x} = prox\ y_{x,¿}¿ - prox\ y_{x,¿¿}$$

$$Dprox\ y_{y} = prox\ y_{y,¿}¿ - prox\ y_{y,¿¿}$$

A greater difference should contribute to a higher likelihood of choosing the option on the right side. Importantly, proximity on the x- and the y-axis should both contribute to the choice (i.e., have positive coefficients).

In pilot M2V1, we fitted generalized linear models to the choices separately for each participant: Choose Right ~ Dproxy_x + Dproxy_y + Trial. All independent variables were continuous and were standardised before estimating the model. In Figure S4-4, we can see that one of the non-generalizers showed an exceptionally negative coefficient for $Dprox\,y_y$, indicating that this non-generalizer did not show consistent behaviour in the 2AFC task and the treasure hunt task. In fact, this non-generalizer showed high accuracy in the 2AFC task (Figure 4-5C in the main text).

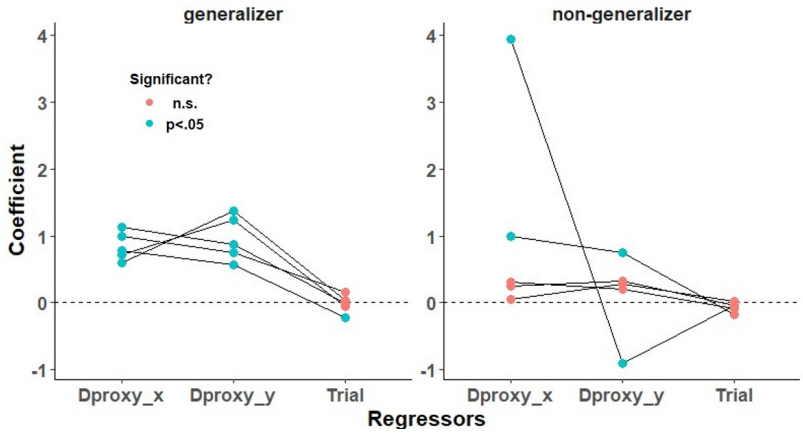


Figure S4-4. Behavioural consistency across tasks in pilot M2V1. Consistency was measured by how well the treasure hunt task response predicted the 2AFC task choice. Dots represented regression coefficients of individual participants.

In pilot M2V2, we fitted a generalized linear mixed-effect model using a binomial link function to participants' choices. We included random slopes and random intercepts for participants:

Choose Right ~ Dproxy_x + Dproxy_y + Trial + (1 + Dproxy_x + Dproxy_y + Trial | participant)

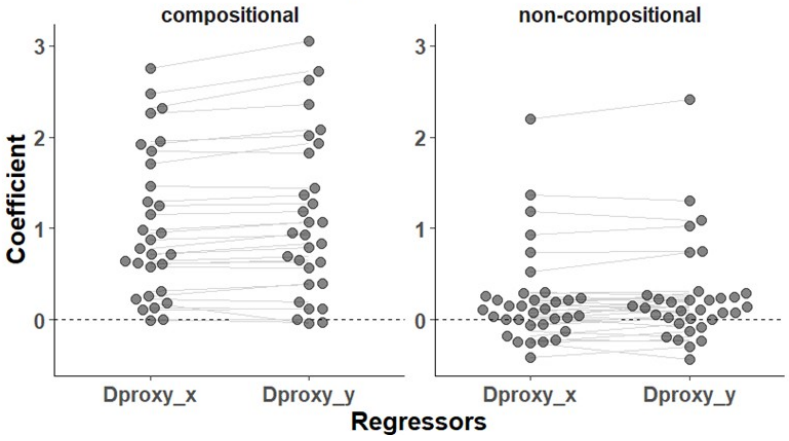
We set non-compositional generalizer as the reference level for the participant group variable. Results of the mixed effect model was

reported in table S4-5, we further extracted the random slopes for Dproxy_x and Dproxy_y and plotted them in Figure S4-5.

Table S4-5 Results of mixed effect model fitted to trial-wise accuracy and RT in 2AFC task in pilot M2V2-V3.

Regressors	Coefficients [C.I.]
(Intercept)	0.079 **
	[0.025, 0.133]
Dproxy_x	0.615 ***
	[0.415, 0.815]
Dproxy_y	0.664 ***
	[0.450, 0.878]
Trial	0.008
	[-0.022, 0.038]
N (trial)	11613
N (participant)	32
AIC	13339.887
BIC	13442.925
R2 (fixed)	0.206
R2 (total)	0.405

*** p < 0.001; ** p < 0.01; * p < 0.05.



F**igure S4-5. Behavioural consistency across tasks in pilot M2V2-V3. Consistency was measured by how well the treasure hunt task response predicted the 2AFC task choice. Dots represented the random slopes for individual participants extracted from the linear mixed-effects model.**

B.6 Transfer learning effect

Participants' response error was measured by the distance to the correct location. In Figure S4-6, we plotted the error curve of individual participants as well as the group average when learning different training maps (T1 and T2). We observed that, on average, the initial error of T2 seemed to be smaller than T1, and the error seemed to decrease faster in T2 than in T1. To quantify this, we fitted an exponential decay model to the pretraining error data in pilot M2V2-V3 and performed statistical analyses to confirm this observation.

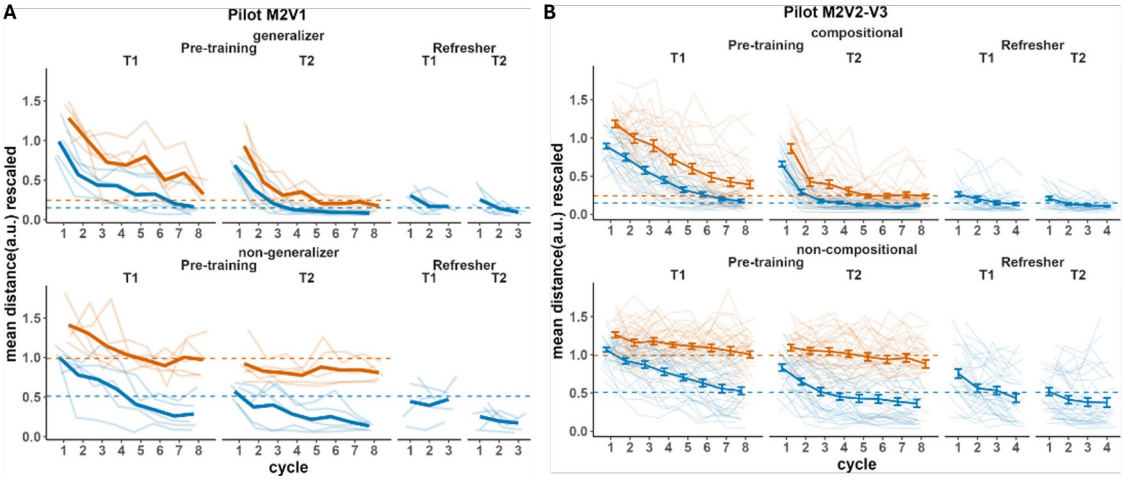


Figure S4-6. Pretraining performance in the multi-map experiments. (A) Data from pilot M2V1. (B) Data from pilot M2V2-V3. Translucent lines represented individual participants. Solid lines represented the group average. Error bars represented mean ± standard error.

The exponential decay function used to fit the data was:

$$error_t = N_0 e^{-\lambda * t}$$

$t$indicated the cycle, and$error_t$ denoted the error of the current trial, N_0 was the initial error, and λ was the decay rate. Parameters were estimated by minimising the mean squared error. To make sure the parameters were interpretable, we imposed two constraints such that

both N_0 and λ had to be positive. Data from training stimuli and validation stimuli from different training maps were fitted separately. Therefore, we obtained four sets of parameters for each participant.

We measure goodness of fit with R-squared. We generated data using the fitted parameters and computed the R-squared values between the generated data and the observed data. The fitted parameters captured the data quite well (Table S4-6).

Table S4-6 Goodness of fit of the exponential decay model measured by R-squared in pilot M2V2-V3 pretraining data.

Participant Group	Map Type-Stimuli Type	R^2 (S.E.M)
non-compositional	T1-Training	0.973±0.009
	T2-Training	0.914±0.012
	T1-Validation	0.995±0.006
	T2-Validation	0.985±0.008
compositional	T1-Training	0.970±0.010
	T2-Training	0.854±0.011
	T1-Validation	0.975±0.014
	T2-Validation	0.869±0.015

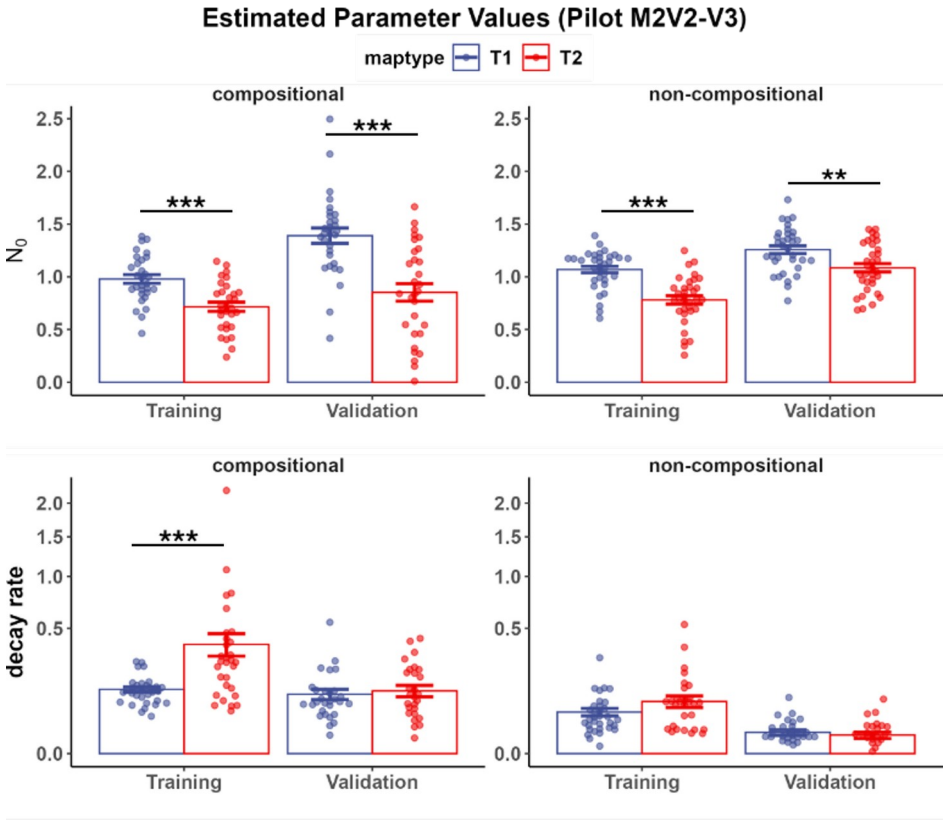
We fitted linear mixed effect models to the parameters to formally test how the learning dynamics in different stimulus types were affected by Participant Group (G vs nG) and Training Map (T1 vs T2). Random intercepts were included. We set non-compositional generalizer (nG) as the reference level for the participant group variable. We set the first training map (T1) as the reference level for training map variable. Results revealed that there was a significant interaction effect in the decay rate of training stimuli.

Table S4-7 Effect of Participant Group (G vs nG) ×Training Map (T1 vs T2) on the learning dynamics (initial error N_0 and decay rate λ) of different stimulus groups (training and validation).

	Training Stimuli		Validation stimuli	
	initial error N_0	decay rate λ	initial error N_0	decay rate λ
G>nG	-0.090	0.076	0.133	0.098 ***
	[-0.197, 0.017]	[-0.031, 0.182]	[-0.031, 0.297]	[0.058, 0.138]
T2>T1	-0.288 ***	0.032	-0.171 *	-0.003
	[-0.392, -0.185]	[-0.063, 0.126]	[-0.330, -0.013]	[-0.042, 0.035]
(G>nG)×(T2>T1)	0.024	0.218 **	-0.367 **	0.016
	[-0.127, 0.175]	[0.080, 0.356]	[-0.599, -0.135]	[-0.040, 0.073]
N (trial)	128	128	128	128
N (participant)	64	64	64	64
AIC	-0.353	-2.880	105.875	-244.663
BIC	16.759	14.232	122.987	-227.550
R2 (fixed)	0.306	0.254	0.258	0.302
R2 (total)	0.306	0.374	0.258	0.302

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.
G: compositional generalizers; NG: non-compositional generalizers
Square brackets indicated 95% confidence intervals

To further reveal the direction of effects, we performed paired t-tests (Figure S4-7). Results showed that while both compositional and non-compositional participants showed smaller initial errors when learning the second training map (T2), only compositional generalizers showed a faster decay rate when learning T2. Interestingly, the decay rate for validation stimuli did not differ in T1 and T2 in compositional generalizers. This indicated that, while they learned the rules faster in T2, the application of rules to the validation stimuli was of similar



speed in T1 and T2.

Figure S4-7. Parameter estimates of the exponential decay model fitted to pilot M2V2-V3 pretraining data. Top panels: initial error N_0. Bottom Panels: decay rate λ. Dots indicated individual participants. Bars and error bars indicated S.E.M. Asterisks indicated results of paired t-tests (Bonferroni-corrected): *p<0.05, **p<0.01, ***p<0.001.

Appendix C: Supplementary Materials for Chapter 5

Table S5-1 Counterbalanced allocation of feature-dimension mapping

A. feature-dimension mapping permutations

	feature-dimension mapping					
dimension	1	2	3	4	5	6
constraining	colour	colour	shape	shape	texture	texture
relevant	shape	texture	colour	texture	colour	shape
irrelevant	texture	shape	texture	colour	shape	colour

B. the number of participants in each counterbalance condition

	feature-dimension mapping						
condition	1	2	3	4	5	6	sum
constraining	4	4	4	4	5	7	28
relevant	4	5	8	6	4	6	33
irrelevant	5	5	5	4	4	6	29

Table S5-2. Effect of training curriculum and time on accuracy during training and transfer

A. Linear Mixed effect model on accuracy

Regressors	train	transfer	transfer (anchor)	transfer (test)
(Intercept)	0.22 *** [0.13, 0.32]	0.25 *** [0.15, 0.34]	0.27 *** [0.17, 0.37]	0.22 *** [0.11, 0.32]
# block	0.04 *** [0.03, 0.04]	0.02 *** [0.02, 0.03]	0.03 *** [0.02, 0.03]	0.02 *** [0.01, 0.02]
Constrain vs Irrelevant	0.31 *** [0.17, 0.44]	0.02 [-0.12, 0.16]	0.05 [-0.10, 0.19]	-0.01 [-0.16, 0.13]
Relevant vs Irrelevant	0.32 *** [0.19, 0.45]	0.10 [-0.04, 0.23]	0.15 * [0.01, 0.29]	0.03 [-0.11, 0.17]
# block * (C vs I)	-0.01 ** [-0.02, -0.00]	0.02 *** [0.01, 0.02]	0.01 ** [0.00, 0.02]	0.02 *** [0.01, 0.03]
# block * (R vs I)	-0.02 *** [-0.02, -0.01]	0.00 [-0.00, 0.01]	-0.00 [-0.01, 0.01]	0.01 * [0.00, 0.02]
N (trial)	900	900	900	900
N (participant)	90	90	90	90
AIC	-690.58	-661.98	-401.75	-297.56
BIC	-652.16	-623.56	-363.33	-259.14
R2 (fixed)	0.20	0.11	0.12	0.07
R2 (total)	0.81	0.79	0.76	0.72

*** p < 0.001; ** p < 0.01; * p < 0.05.

B. Post-hoc test of the main effect of between training curricula of treatment groups

Null Hypothesis	Block type	Contrast Estimate	std. error	Z statistic	p
C – R = 0	train	-0.009	0.067	-0.137	0.891
	transfer	-0.076	0.068	-1.131	0.516

	transfer (anchor)	-0.105	0.072	-1.455	0.516
	transfer (test)	-0.041	0.073	-0.564	0.764

*** p < 0.001; ** p < 0.01; * p < 0.05, FDR-corrected

C. Post-hoc test of the slope of learning between training curricula of treatment groups

Null Hypothesis	Block type	Contrast Estimate	std. error	Z statistic	p
C * #block – R * #block = 0	train	0.005	0.004	1.303	0.192
	transfer	0.011	0.004	2.936	0.007**
	transfer (anchor)	0.013	0.005	2.924	0.007**
	transfer (test)	0.009	0.005	1.875	0.081

*** p < 0.001; ** p < 0.01; * p < 0.05, FDR-corrected

D. accuracy of test trials in each transfer block and comparison to chance level

block number (#block)	constraining	relevant	irrelevant
1	0.223±0.154	0.242±0.175*	0.207±0.255
2	0.223±0.241	0.301±0.229**	0.27±0.262*
3	0.327±0.277**	0.306±0.229**	0.282±0.282*
4	0.372±0.327**	0.379±0.28***	0.296±0.247**
5	0.432±0.329***	0.384±0.303***	0.293±0.266*
6	0.423±0.383***	0.437±0.363***	0.287±0.285*
7	0.443±0.387***	0.447±0.353***	0.325±0.271**
8	0.47±0.371***	0.452±0.366***	0.368±0.333**
9	0.542±0.374***	0.434±0.364***	0.376±0.303***

10	0.518±0.425***	0.508±0.333***	0.368±0.363***

One sample t-test against chance level of 1/6: *** p < 0.001; ** p < 0.01; * p < 0.05, FDR-corrected

Table S5-3. Structure learning and generalization across time

A. Linear Mixed effect model on the participants' RDMs of transfer block

Coefficient	Estimate [95% CI]
(Intercept)	0.147 *** [0.142, 0.153]
condition C-I	-0.010 *** [-0.014, -0.005]
condition R-I	-0.032 *** [-0.037, -0.027]
block number	-0.007 *** [-0.007, -0.006]
$RDM_{ground\text{-}truth}$	0.033 *** [0.015, 0.051]
$RDM_{constrain}$	0.079 *** [0.065, 0.092]
$RDM_{Relevant}$	0.014 ** [0.004, 0.024]
$RDM_{irrelevant}$	-0.001 [-0.011, 0.009]
condition C-I * $RDM_{ground\text{-}truth}$	-0.017 [-0.041, 0.008]
condition R-I * $RDM_{ground\text{-}truth}$	0.030 * [0.006, 0.055]
condition C-I * $RDM_{constrain}$	-0.013 [-0.030, 0.004]
condition R-I * $RDM_{constrain}$	0.071 *** [0.054, 0.089]
condition C-I * $RDM_{Relevant}$	0.030 *** [0.017, 0.044]
condition R-I * $RDM_{Relevant}$	0.006 [-0.007, 0.019]
condition C-I * $RDM_{irrelevant}$	0.006 [-0.007, 0.018]
condition R-I * $RDM_{irrelevant}$	-0.000 [-0.013, 0.013]
block number * $RDM_{ground\text{-}truth}$	0.023 *** [0.020, 0.025]
block number * $RDM_{constrain}$	0.004 *** [0.002, 0.006]
block number * $RDM_{Relevant}$	0.003 *** [0.002, 0.005]
block number * $RDM_{irrelevant}$	0.000 [-0.001, 0.002]
condition C-I * # block * $RDM_{ground\text{-}truth}$	0.025 *** [0.021, 0.028]
condition R-I * # block * $RDM_{ground\text{-}truth}$	0.006 ** [0.002, 0.010]
condition C-I * # block * $RDM_{constrain}$	-0.004 ** [-0.007, -0.002]
condition R-I * # block * $RDM_{constrain}$	0.001 [-0.002, 0.004]
condition C-I * # block * $RDM_{Relevant}$	-0.003 ** [-0.005, -0.001]
condition R-I * # block * $RDM_{Relevant}$	-0.002 [-0.004, 0.000]
condition C-I * # block * $RDM_{irrelevant}$	-0.001 [-0.003, 0.001]
condition R-I * # block * $RDM_{irrelevant}$	-0.000 [-0.002, 0.002]
N	10530

N (pairs)	351
AIC	-27647.347
R2 (fixed)	0.804
R2 (total)	0.812

B. Comparing ground-truth RDM and constraining RDM regressor coefficient estimates in transfer blocks in different conditions.

condition	Block number	Contrast Estimate	std. error	statistic	p.fdr
constraining	1	0.000	0.025	0.001	1.000
	2	0.015	0.024	0.614	0.568
	3	0.083	0.022	3.746	<0.001***
	4	0.187	0.022	8.492	<0.001***
	5	0.208	0.022	9.623	<0.001***
	6	0.269	0.020	13.318	<0.001***
	7	0.288	0.019	15.242	<0.001***
	8	0.349	0.019	18.422	<0.001***
	9	0.338	0.018	18.897	<0.001***
	10	0.384	0.020	19.514	<0.001***
relevant	1	-0.106	0.022	-4.925	<0.001***
	2	-0.063	0.023	-2.730	0.009**
	3	-0.052	0.021	-2.442	0.019*
	4	0.047	0.020	2.356	0.023*
	5	0.102	0.020	5.167	<0.001***
	6	0.097	0.019	5.131	<0.001***
	7	0.079	0.020	3.901	<0.001***
	8	0.109	0.019	5.624	<0.001***
	9	0.096	0.017	5.644	<0.001***
	10	0.130	0.019	6.732	<0.001***
irrelevant	1	-0.016	0.025	-0.663	0.544
	2	-0.023	0.020	-1.145	0.281
	3	-0.029	0.021	-1.367	0.199
	4	0.062	0.023	2.758	0.008**
	5	0.028	0.021	1.320	0.212
	6	0.068	0.021	3.283	0.002**
	7	0.074	0.019	3.852	<0.001***
	8	0.142	0.020	7.280	<0.001***
	9	0.148	0.020	7.445	<0.001***

	10	0.110	0.021	5.353	<0.001***

*** p < 0.001; ** p < 0.01; * p < 0.05, FDR-corrected

References

Abel, R., Brunmair, M., & Weissgerber, S. C. (2021). Change one category at a time: Sequence effects beyond interleaving and blocking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(7), 1083–1105. https://doi.org/10.1037/xlm0001003

Ahlheim, C., & Love, B. C. (2018). Estimating the functional dimensionality of neural representations. *NeuroImage*, *179*, 51–62. https://doi.org/10.1016/j.neuroimage.2018.06.015

Altan, E., Solla, S. A., Miller, L. E., & Perreault, E. J. (2021). Estimating the dimensionality of the manifold underlying multi-electrode neural recordings. *PLOS Computational Biology*, *17*(11), e1008591. https://doi.org/10.1371/journal.pcbi.1008591

Amo, R., Matias, S., Yamanaka, A., Tanaka, K. F., Uchida, N., & Watabe-Uchida, M. (2022). A gradual temporal shift of dopamine responses mirrors the progression of temporal difference error in machine learning. *Nature Neuroscience*, *25*(8), 1082–1092. https://doi.org/10.1038/s41593-022-01109-2

Andersen, R. A., Essick, G. K., & Siegel, R. M. (1985). Encoding of spatial location by posterior parietal neurons. *Science*, *230*(4724), 456–458. https://doi.org/10.1126/science.4048942

Andersen, R. A., Snyder, L. H., Bradley, D. C., & Xing, J. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience*, *20*(1), 303–330. https://doi.org/10.1146/annurev.neuro.20.1.303

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53. https://doi.org/10.1037/0278-7393.14.1.33

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*(1), 149–178. https://doi.org/10.1146/annurev.psych.56.091103.070217

Baden, T., Euler, T., & Berens, P. (2020). Understanding the retinal basis of vision across species. *Nature Reviews Neuroscience*, *21*(1), 5–20. https://doi.org/10.1038/s41583-019-0242-1

Bakermans, J. J. W., Warren, J., Whittington, J. C. R., & Behrens, T. E. J. (2025). Constructing future behavior in the hippocampal formation through composition and replay. *Nature Neuroscience*. https://doi.org/10.1038/s41593-025-01908-3

Balleine, B. W., Daw, N. D., & O'Doherty, J. P. (2009). Multiple forms of value learning and the function of dopamine. In

Neuroeconomics (pp. 367–387). Elsevier. https://doi.org/10.1016/B978-0-12-374176-9.00024-5

Banker, L., & Tadi, P. (2025). Neuroanatomy, precentral gyrus. In *Statpearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK544218/

Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience*, *22*(6), 359–371. https://doi.org/10.1038/s41583-021-00448-6

Baram, A. B., Muller, T. H., Nili, H., Garvert, M. M., & Behrens, T. E. J. (2021). Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. *Neuron*, *109*(4), 713-723.e7. https://doi.org/10.1016/j.neuron.2020.11.024

Baram, A., Nili, H., Barreiros, I., Samborska, V., Behrens, T. E. J., & Garvert, M. M. (2024). *An abstract relational map emerges in the human medial prefrontal cortex with consolidation*. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2024.10.11.617652

Barlow, H. B. (2012). Possible principles underlying the transformations of sensory messages. In *Sensory Communication*. The MIT Press. https://doi.org/10.7551/mitpress/9780262518420.003.0013

Baron, S. G., & Osherson, D. (2011). Evidence for conceptual combination in the left anterior temporal lobe. *NeuroImage*, *55*(4), 1847–1852. https://doi.org/10.1016/j.neuroimage.2011.01.066

Baron, S. G., Thompson-Schill, S. L., Weber, M., & Osherson, D. (2010). An early stage of conceptual combination: Superimposition of constituent concepts in left anterolateral temporal lobe. *Cognitive Neuroscience*, *1*(1), 44–51. https://doi.org/10.1080/17588920903548751

Barron, H. C., Dolan, R. J., & Behrens, T. E. J. (2013). Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature Neuroscience*, *16*(10), 1492–1498. https://doi.org/10.1038/nn.3515

Barron, H. C., Garvert, M. M., & Behrens, T. E. J. (2016). Repetition suppression: A means to index neural representations using BOLD? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1705), 20150355. https://doi.org/10.1098/rstb.2015.0355

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660. https://doi.org/10.1017/s0140525x99002149

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Others. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*, 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Bemis, D. K., & Pylkkänen, L. (2011). Simple Composition: A Magnetoencephalography Investigation into the Comprehension

of Minimal Linguistic Phrases. *The Journal of Neuroscience*, *31*(8), 2801–2814. https://doi.org/10.1523/JNEUROSCI.5003-10.2011

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

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.e21. https://doi.org/10.1016/j.cell.2020.09.031

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147. https://doi.org/10.1037/0033-295x.94.2.115

Bongioanni, A., Folloni, D., Verhagen, L., Sallet, J., Klein-Flügge, M. C., & Rushworth, M. F. S. (2021). Activation and disruption of a neural mechanism for novel choice in monkeys. *Nature*, *591*, 270–274. https://doi.org/10.1038/s41586-020-03115-5

Bonner, M. F., & Epstein, R. A. (2021). Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-24368-2

Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 38–50. https://doi.org/10.1037/0278-7393.30.1.38

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences*, *23*(5), 408–422. https://doi.org/10.1016/j.tics.2019.02.006

Bouchacourt, F., Palminteri, S., Koechlin, E., & Ostojic, S. (2020). Temporal chunking as a mechanism for unsupervised learning of task-sets. *eLife*, *9*, e50469. https://doi.org/10.7554/eLife.50469

Bourlard, H., & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, *59*(4–5), 291–294. https://doi.org/10.1007/BF00332918

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27. https://doi.org/10.1016/0030-5073(74)90002-6

Brincat, S. L., & Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience*, *7*(8), 880–886. https://doi.org/10.1038/nn1278

Bröker, F., Love, B. C., & Dayan, P. (2022). When unsupervised training benefits category learning. *Cognition*, *221*, 104984. https://doi.org/10.1016/j.cognition.2021.104984

Buneo, C. A., Jarvis, M. R., Batista, A. P., & Andersen, R. A. (2002). Direct visuomotor transformations for reaching. *Nature*, *416*(6881), 632–636. https://doi.org/10.1038/416632a

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the raven progressive matrices test. *Psychological Review*, *97*(3), 404–431. https://doi.org/10.1037/0033-295X.97.3.404

Carroll, J. D. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *ETS Research Bulletin Series*, *1963*(2). https://doi.org/10.1002/j.2333-8504.1963.tb00958.x

Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*(3), 481–495. https://doi.org/10.3758/s13421-013-0371-0

Carvalho, P. F., & Goldstone, R. L. (2015). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00505

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1699–1719. https://doi.org/10.1037/xlm0000406

Chang, K. K., Cherkassky, V. L., Mitchell, T. M., & Just, M. A. (2009). Quantitative modeling of the neural representation of adjective-noun phrases to account for fMRI activation. In K.-Y. Su, J. Su, J. Wiebe, & H. Li (Eds.), *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 638–646). Association for Computational Linguistics. https://aclanthology.org/P09-1072/

Chang, Y.-T., Finkel, E. A., Xu, D., & O'Connor, D. H. (2024). Rule-based modulation of a sensorimotor transformation across cortical areas. *Elife*, *12*. https://doi.org/10.7554/elife.92620.3

Chomsky, N. (1981). Knowledge of language: Its elements and origins. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *295*(1077), 223–234. https://doi.org/10.1098/rstb.1981.0135

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use* (D. Byrne & M. Kölbel, Eds.). Prager.

Chou, C.-N., Arend, L., Wakhloo, A. J., Kim, R., Slatton, W., & Chung, S. (2024). *Neural Manifold Capacity Captures Representation Geometry, Correlations, and Task-Efficiency Across Species and Behaviors*. https://doi.org/10.1101/2024.02.26.582157

Christophel, T. B., Hebart, M. N., & Haynes, J.-D. (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. *Journal of Neuroscience*, *32*(38), 12983–12989. https://doi.org/10.1523/JNEUROSCI.0184-12.2012

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J.-D. (2017). The distributed nature of working memory. *Trends in Cognitive Sciences*, *21*(2), 111–124. https://doi.org/10.1016/j.tics.2016.12.007

Chung, S., & Abbott, L. F. (2021). Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*, *70*, 137–144. https://doi.org/10.1016/j.conb.2021.10.010

Cohen, Y. E., & Andersen, R. A. (2002). A common reference frame for movement plans in the posterior parietal cortex. *Nature Reviews Neuroscience*, *3*(7), 553–562. https://doi.org/10.1038/nrn873

Conway, B. R. (2018). The organization and operation of inferior temporal cortex. *Annual Review of Vision Science*, *4*(1), 381–402. https://doi.org/10.1146/annurev-vision-091517-034202

Courellis, H. S., Minxha, J., Ramirez Cardenas, A., Kimmel, D. L., Reed, C. M., Valiante, T. A., Salzman, C. D., Mamelak, A. N., Adolphs, R., Fusi, S., & Rutishauser, U. (2023). *Abstract representations emerge in human hippocampal neurons during inference behavior* [Preprint]. Neuroscience. https://doi.org/10.1101/2023.11.10.566490

Dekker, R. B., Otto, F., & Summerfield, C. (2022). Curriculum learning for human compositional generalization. *Proceedings of the National Academy of Sciences*, *119*(41), e2205582119. https://doi.org/10.1073/pnas.2205582119

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968–986. https://doi.org/10.1037/0278-7393.23.4.968

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341. https://doi.org/10.1016/j.tics.2007.06.010

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, *13*(4), e1005508. https://doi.org/10.1371/journal.pcbi.1005508

Donato, F., Xu Schwartzlose, A., & Viana Mendes, R. A. (2023). How Do You Build a Cognitive Map? The Development of Circuits and Computations for the Representation of Space in the Brain. *Annual Review of Neuroscience*, *46*(1), 281–299. https://doi.org/10.1146/annurev-neuro-090922-010618

Drieu, C., Zhu, Z., Wang, Z., Fuller, K., Wang, A., Elnozahy, S., & Kuchibhotla, K. (2024). *Rapid emergence of latent knowledge in*

the sensory cortex drives learning. https://doi.org/10.1101/2024.06.10.597946

Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. *Neuron*, *109*(19), 3055–3068. https://doi.org/10.1016/j.neuron.2021.07.011

Erhan, D., Courville, A., Bengio, Y., & Vincent, P. (2010). Why Does Unsupervised Pre-training Help Deep Learning? *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 201–208. https://proceedings.mlr.press/v9/erhan10a.html

Erhan, D., Manzagol, P.-A., Bengio, Y., Bengio, S., & Vincent, P. (2009). The Difficulty of Training Deep Architectures and the Effect of Unsupervised Pre-Training. In D. van Dyk & M. Welling (Eds.), *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics* (Vol. 5, pp. 153–160). PMLR. https://proceedings.mlr.press/v5/erhan09a.html

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140. https://doi.org/10.1037/0096-3445.127.2.107

Fascianelli, V., Battista, A., Stefanini, F., Tsujimoto, S., Genovesio, A., & Fusi, S. (2024). Neural representational geometries reflect behavioral differences in monkeys and recurrent neural networks. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-024-50503-w

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, *110*(7), 1258-1270.e11. https://doi.org/10.1016/j.neuron.2022.01.005

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., & Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage*, *54*(1), 313–327. https://doi.org/10.1016/j.neuroimage.2010.07.033

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *Neuroimage*, *47*, S102. https://doi.org/10.1016/S1053-8119(09)70884-5

Frankland, S. M., & Greene, J. D. (2020). Concepts and Compositionality: In Search of the Brain's Language of Thought. *Annual Review of Psychology*, *71*(1), 273–303. https://doi.org/10.1146/annurev-psych-122216-011829

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Froudarakis, E., Cohen, U., Diamantaki, M., Walker, E. Y., Reimer, J., Berens, P., Sompolinsky, H., & Tolias, A. S. (2020). *Object*

manifold geometry across the mouse cortical visual hierarchy. https://doi.org/10.1101/2020.08.20.258798

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy*. *Cognitive Science*, *7*(2), 155–170. https://doi.org/10.1207/s15516709cog0702_3

Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., & Brea, J. (2018). Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. *Frontiers in Neural Circuits*, *12*, 53. https://doi.org/10.3389/fncir.2018.00053

Ghahramani, Z. (1996, February 29). *One Hidden Layer Linear Networks and Canonical Correlations*.

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*(1), 15–36. https://doi.org/10.1006/anbe.2003.2174

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*(3), 306–355. https://doi.org/10.1016/0010-0285(80)90013-4

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*(1), 1–38. https://doi.org/10.1016/0010-0285(83)90002-6

Gilboa, A., & Marlatte, H. (2017). Neurobiology of schemas and schema-mediated memory. *Trends in Cognitive Sciences*, *21*(8), 618–631. https://doi.org/10.1016/j.tics.2017.04.013

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178. https://doi.org/10.1038/nature18933

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, *3*(4), 491–516. https://doi.org/10.1002/hipo.450030410

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154. https://doi.org/10.1080/03640210701802071

Graziano, M. S. A., Yap, G. S., & Gross, C. G. (1994). Coding of visual space by premotor neurons. *Science*, *266*(5187), 1054–1057. https://doi.org/10.1126/science.7973661

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), 536–548. https://doi.org/10.1038/nrn3747

Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology*, *51*(1), 79–88. https://doi.org/10.1037/h0046219

Haga, T., Oseki, Y., & Fukai, T. (2025). A unified neural representation model for spatial and conceptual computations. *Proceedings of the National Academy of Sciences*, *122*(11), e2413449122. https://doi.org/10.1073/pnas.2413449122

Hall, G., Ray, E., & Bonardi, C. (1993). Acquired equivalence between cues trained with a common antecedent. *Journal of Experimental Psychology: Animal Behavior Processes*, *19*(4), 391–399. https://doi.org/10.1037/0097-7403.19.4.391

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, *56*(1), 51–65. https://doi.org/10.1037/h0062474

Hattori, R., Hedrick, N. G., Jain, A., Chen, S., You, H., Hattori, M., Choi, J.-H., Lim, B. K., Yasuda, R., & Komiyama, T. (2023). Meta-reinforcement learning via orbitofrontal cortex. *Nature Neuroscience*, *26*(12), 2182–2191. https://doi.org/10.1038/s41593-023-01485-3

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Hebb, D. O. (2002). The first stage of perception: Growth of the assembly. In *The organization of behavior: A neuropsychological theory* (pp. 102–120). Taylor & Francis Group. http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=227504

Hennig, J. A., Oby, E. R., Losey, D. M., Batista, A. P., Yu, B. M., & Chase, S. M. (2021). How learning unfolds in the brain: Toward an optimization view. *Neuron*, *109*(23), 3720–3735. https://doi.org/10.1016/j.neuron.2021.09.005

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, *15*(4), 332–340. https://doi.org/10.3758/BF03197035

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, *24*(5), 849–878. https://doi.org/10.1017/S0140525X01000103

Honey, R. C., & Hall, G. (1989). Acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology. Animal Behavior Processes*, *15*(4), 338–346.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Hull, C. L. (1947). The problem of primary stimulus generalization. *Psychological Review*, *54*(3), 120–134. https://doi.org/10.1037/h0061159

Ito, T., Klinger, T., Schultz, D. H., Murray, J. D., Cole, M. W., & Rigotti, M. (2022). *Compositional generalization through abstract representations in human and artificial neural networks* (No. arXiv:2209.07431). arXiv. http://arxiv.org/abs/2209.07431

Jazayeri, M., & Ostojic, S. (2021). Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Current Opinion in Neurobiology*, *70*, 113–120. https://doi.org/10.1016/j.conb.2021.08.002

Johnston, W. J., & Fusi, S. (2023). Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nature Communications*, *14*(1), 1040. https://doi.org/10.1038/s41467-023-36583-0

Jonikaitis, D., Xia, R., & Moore, T. (2025). Robust encoding of stimulus–response mapping by neurons in visual cortex. *Proceedings of the National Academy of Sciences*, *122*(9), e2408079122. https://doi.org/10.1073/pnas.2408079122

Kalish, M. L. (2013). Learning and extrapolating a periodic function. *Memory & Cognition*, *41*(6), 886–896. https://doi.org/10.3758/s13421-013-0306-9

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, *111*(4), 1072–1099. https://doi.org/10.1037/0033-295X.111.4.1072

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692. https://doi.org/10.1073/pnas.0802631105

Kerrén, C., Reznik, D., Doeller, C. F., & Griffiths, B. J. (2025). Exploring the role of dimensionality transformation in episodic memory. *Trends in Cognitive Sciences*, S136466132500021X. https://doi.org/10.1016/j.tics.2025.01.007

Kim, J.-H., Daie, K., & Li, N. (2025). A combinatorial neural code for long-term motor memory. *Nature*, *637*(8046), 663–672. https://doi.org/10.1038/s41586-024-08193-3

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 811–836. https://doi.org/10.1037/0278-7393.17.5.811

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Kriegeskorte, N., & Wei, X.-X. (2021). Neural tuning and representational geometry. *Nature Reviews Neuroscience*, *22*(11), 703–718. https://doi.org/10.1038/s41583-021-00502-3

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. https://doi.org/10.1037/0033-295X.99.1.22

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated. *Trends in Cognitive Sciences*, *20*(7), 512–534. https://doi.org/10.1016/j.tics.2016.05.004

Kung, S.-Y., Diamantaras, K. I., & Taur, J.-S. (1994). Adaptive principal component EXtraction (APEX) and applications. *IEEE Transactions on Signal Processing*, *42*(5), 1202–1217. https://doi.org/10.1109/78.295198

Kwantes, P. J., & Neal, A. (2006). Why people underestimate y when extrapolating in linear functions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1019–1030. https://doi.org/10.1037/0278-7393.32.5.1019

Lafer-Sousa, R., & Conway, B. R. (2013). Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. *Nature Neuroscience*, *16*(12), 1870–1878. https://doi.org/10.1038/nn.3555

Lakoff, G., & Johnson, M. (2003). *Metaphors we live by* (6. print). University of Chicago Press.

Lashley, K. S., & Wade, M. (1946). The pavlovian theory of generalization. *Psychological Review*, *53*(2), 72–87. https://doi.org/10.1037/h0059999

LaTourrette, A., & Waxman, S. R. (2022). Sparse labels, no problems: Infant categorization under challenging conditions. *Child Development*, *93*(6), 1903–1911. https://doi.org/10.1111/cdev.13818

Laufer, O., & Paz, R. (2012). Monetary loss alters perceptual thresholds and compromises future decisions via amygdala and prefrontal networks. *Journal of Neuroscience*, *32*(18), 6304–6311. https://doi.org/10.1523/JNEUROSCI.6281-11.2012

Li, N., & DiCarlo, J. J. (2008). Unsupervised Natural Experience Rapidly Alters Invariant Object Representation in Visual Cortex. *Science*, *321*(5895), 1502–1507. https://doi.org/10.1126/science.1160028

Li, N., & DiCarlo, J. J. (2010). Unsupervised Natural Visual Experience Rapidly Reshapes Size-Invariant Object Representation in Inferior Temporal Cortex. *Neuron*, *67*(6), 1062–1075. https://doi.org/10.1016/j.neuron.2010.08.029

Li, N., & DiCarlo, J. J. (2012). Neuronal Learning of Invariant Object Representation in the Ventral Visual Stream Is Not Dependent on Reward. *The Journal of Neuroscience*, *32*(19), 6611–6620. https://doi.org/10.1523/JNEUROSCI.3786-11.2012

Liu, H., Agam, Y., Madsen, J. R., & Kreiman, G. (2009). Timing, timing, timing: Fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron*, *62*(2), 281–290. https://doi.org/10.1016/j.neuron.2009.02.025

Livingstone, M., & Hubel, D. (1988). Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science*, *240*(4853), 740–749. https://doi.org/10.1126/science.3283936

Lovett, A., & Forbus, K. (2017). Modeling visual problem solving as analogical reasoning. *Psychological Review*, *124*(1), 60–90. https://doi.org/10.1037/rev0000039

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215. https://doi.org/10.3758/s13423-015-0808-5

Luettgau, L., Erdmann, T., Veselic, S., Stachenfeld, K. L., Kurth-Nelson, Z., Moran, R., & Dolan, R. J. (2024). Decomposing dynamical subprocesses for compositional generalization. *Proceedings of the National Academy of Sciences*, *121*(46), e2408134121. https://doi.org/10.1073/pnas.2408134121

Magee, J. C., & Grienberger, C. (2020). Synaptic Plasticity Forms and Functions. *Annual Review of Neuroscience*, *43*(1), 95–117. https://doi.org/10.1146/annurev-neuro-090919-022842

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. *Nature Neuroscience*, *22*(7), 1159–1167. https://doi.org/10.1038/s41593-019-0414-3

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295x.102.3.419

Mcdaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review*, *12*(1), 24–42. https://doi.org/10.3758/BF03196347

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254–278. https://doi.org/10.1037/0033-295X.100.2.254

Mednick, S. A., & Freedman, J. L. (1960). Stimulus generalization. *Psychological Bulletin*, *57*(3), 169–200. https://doi.org/10.1037/h0041650

Meeter, M., Shohamy, D., & Myers, C. E. (2009). Acquired equivalence changes stimulus representations. *Journal of the Experimental Analysis of Behavior*, *91*(1), 127–141. https://doi.org/10.1901/jeab.2009.91-127

Meunier, D., Lambiotte, R., & Bullmore, E. T. (2010). Modular and hierarchically modular organization of brain networks. *Frontiers in Neuroscience*, *4*. https://doi.org/10.3389/fnins.2010.00200

Mitchell, J., & Lapata, M. (2008). Vector-based models of semantic composition. In J. D. Moore, S. Teufel, J. Allan, & S. Furui (Eds.), *Proceedings of ACL-08: HLT* (pp. 236–244). Association for Computational Linguistics. https://aclanthology.org/P08-1028/

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, *34*(8), 1388–1429. https://doi.org/10.1111/j.1551-6709.2010.01106.x

Mitchell, M. (2021). Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, *1505*(1), 79–101. https://doi.org/10.1111/nyas.14619

Moskvichev, A., Odouard, V. V., & Mitchell, M. (2023). The ConceptARC Benchmark: Evaluating Understanding and Generalization in the ARC Domain. *Trans. Mach. Learn. Res.* https://doi.org/10.48550/arXiv.2305.07141

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, *59*(3), 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316. https://doi.org/10.1037/0033-295X.92.3.289

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*(10), 1544–1553. https://doi.org/10.1038/s41593-019-0470-8

Norbury, A., Robbins, T. W., & Seymour, B. (2018). Value generalization in human avoidance learning. *Elife*, *7*, e34779. https://doi.org/10.7554/eLife.34779

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037//0096-3445.115.1.39

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, *43*(1), 25–53. https://doi.org/10.1146/annurev.ps.43.020192.000325

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 211–233. https://doi.org/10.1037/0278-7393.18.2.211

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning

in the human brain. *Neuron*, *38*(2), 329–337. https://doi.org/10.1016/S0896-6273(03)00169-7

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*(3), 267–273. https://doi.org/10.1007/BF00275687

Oja, E. (1992). Principal components, minor components, and linear neural networks. *Neural Networks*, *5*(6), 927–935. https://doi.org/10.1016/S0893-6080(05)80089-9

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*(1), 171–175. https://doi.org/10.1016/0006-8993(71)90358-1

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609. https://doi.org/10.1038/381607a0

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*(1), 35–58. https://doi.org/10.1016/0010-0277(81)90013-5

Pagan, M., Tang, V. D., Aoi, M. C., Pillow, J. W., Mante, V., Sussillo, D., & Brody, C. D. (2025). Individual variability of neural computations underlying flexible decisions. *Nature*, *639*(8054), 421–429. https://doi.org/10.1038/s41586-024-08433-6

Park, J., Holmes, C. D., & Snyder, L. H. (2025). *Compositional architecture: Orthogonal neural codes for task context and spatial memory in prefrontal cortex*. Neuroscience. https://doi.org/10.1101/2025.02.25.640211

Pavlov, I. P. (1906). The scientific investigation of the psychical faculties or processes in the higher animals. *Science*, *24*(620), 613–619. https://doi.org/10.1126/science.24.620.613

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–238. https://doi.org/10.1016/j.tics.2006.03.006

Pesnot Lerousseau, J., & Summerfield, C. (2024). Space as a scaffold for rotational generalisation of abstract concepts. *Elife*, *13*, RP93636. https://doi.org/10.7554/eLife.93636.3

Petre, B., Ceko, M., Friedman, N. P., C. Keller, M., A. Lindquist, M., A. Reynolds Losin, E., & Wager, T. (2023). *2016 glasser MMP1.0 cortical atlases* (p. 534271010 Bytes) [Dataset]. figshare. https://doi.org/10.6084/M9.FIGSHARE.24431146.V8

Pettine, W. W., Raman, D. V., Redish, A. D., & Murray, J. D. (2023). Human generalization of internal representations through prototype learning with goal-directed attention. *Nature Human*

Behaviour, *7*(3), 442–463. https://doi.org/10.1038/s41562-023-01543-7

Pothos, E. M. (2005). The rules versus similarity distinction. *Behavioral and Brain Sciences*, *28*(1), 1–14. https://doi.org/10.1017/S0140525X05000014

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018

Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics. *Science*, *366*(6461), 62–66. https://doi.org/10.1126/science.aax0050

Pylkkänen, L., Bemis, D. K., & Blanco Elorrieta, E. (2014). Building phrases in language production: An MEG study of simple composition. *Cognition*, *133*(2), 371–384. https://doi.org/10.1016/j.cognition.2014.07.001

Qian, L., Burrell, M., Hennig, J. A., Matias, S., Murthy, V. N., Gershman, S. J., & Uchida, N. (2025). Prospective contingency explains behavior and dopamine signals during associative learning. *Nature Neuroscience*, *28*(6), 1280–1292. https://doi.org/10.1038/s41593-025-01915-4

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102–1107. https://doi.org/10.1038/nature03687

Radulescu, A., Shin, Y. S., & Niv, Y. (2021). Human Representation Learning. *Annual Review of Neuroscience*, *44*(1), 253–273. https://doi.org/10.1146/annurev-neuro-092920-120559

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855–863. https://doi.org/10.1016/S0022-5371(67)80149-X

Rescorla, R. A. (1976). Stimulus generalization: Some predictions from a model of pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *2*(1), 88–96. https://doi.org/10.1037/0097-7403.2.1.88

Rescorla, R., & Wagner, A. (1972). A theory of pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. *Classical Conditioning: Current Research and Theory*.

Reverberi, C., Görgen, K., & Haynes, J.-D. (2012). Compositionality of rule representations in human prefrontal cortex. *Cerebral Cortex*, *22*(6), 1237–1246. https://doi.org/10.1093/cercor/bhr200

Richards, B. A., Xia, F., Santoro, A., Husse, J., Woodin, M. A., Josselyn, S. A., & Frankland, P. W. (2014). Patterns across multiple memories are identified over time. *Nature Neuroscience*, *17*(7), 981–986. https://doi.org/10.1038/nn.3736

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. https://doi.org/10.1038/nature12160

Rigotti, M., Ben Dayan Rubin, D., Morrison, S. E., Salzman, C. D., & Fusi, S. (2010). Attractor concretion as a mechanism for the formation of context representations. *NeuroImage*, *52*(3), 833–847. https://doi.org/10.1016/j.neuroimage.2010.01.047

Ritchie, J. B., Lee Masson, H., Bracci, S., & Op De Beeck, H. P. (2021). The unreliable influence of multivariate noise normalization on the reliability of neural dissimilarity. *Neuroimage*, *245*, 118686. https://doi.org/10.1016/j.neuroimage.2021.118686

Rolls, E. T., Huang, C.-C., Lin, C.-P., Feng, J., & Joliot, M. (2020). Automated anatomical labelling atlas 3. *NeuroImage*, *206*, 116189. https://doi.org/10.1016/j.neuroimage.2019.116189

Salinas, E. (2004). Fast remapping of sensory stimuli onto motor actions on the basis of contextual modulation. *Journal of Neuroscience*, *24*(5), 1113–1118. https://doi.org/10.1523/JNEUROSCI.4569-03.2004

Samborska, V., Butler, J. L., Walton, M. E., Behrens, T. E. J., & Akam, T. (2022). Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems. *Nature Neuroscience*, *25*(10), 1314–1326. https://doi.org/10.1038/s41593-022-01149-8

Santolin, C., & Saffran, J. R. (2018). Constraints on statistical learning across species. *Trends in Cognitive Sciences*, *22*(1), 52–63. https://doi.org/10.1016/j.tics.2017.10.003

Saxena, S., & Cunningham, J. P. (2019). Towards the neural population doctrine. *Current Opinion in Neurobiology*, *55*, 103–111. https://doi.org/10.1016/j.conb.2019.02.002

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology*, *22*(17), 1622–1627. https://doi.org/10.1016/j.cub.2012.06.056

Schneider, S., Lee, J. H., & Mathis, M. W. (2023). Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, *617*(7960), 360–368. https://doi.org/10.1038/s41586-023-06031-6

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., Botvinick, M., Kurth-Nelson, Z., & Behrens, T. (2023). Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. *Cell*, *186*(22), 4885-4897.e14. https://doi.org/10.1016/j.cell.2023.09.004

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., Friston, K. J., & Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, *429*(6992), 664–667. https://doi.org/10.1038/nature02581

Seymour, K., Clifford, C. W. G., Logothetis, N. K., & Bartels, A. (2010). Coding and binding of color and form in visual cortex. *Cerebral Cortex*, *20*(8), 1946–1954. https://doi.org/10.1093/cercor/bhp265

Sharan, L., Rosenholtz, R., & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of Vision*, *14*(9), 12–12. https://doi.org/10.1167/14.9.12

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398. https://doi.org/10.1126/science.210.4468.390

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Shepard, R. N., & Chang, J.-J. (1963). Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology*, *65*(1), 94–102. https://doi.org/10.1037/h0043732

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193

Sincich, L. C., & Horton, J. C. (2005). THE CIRCUITRY OF V1 AND V2: Integration of color, form, and motion. *Annual Review of Neuroscience*, *28*(1), 303–326. https://doi.org/10.1146/annurev.neuro.28.061604.135731

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences*, *7*(6), 246–251. https://doi.org/10.1016/S1364-6613(03)00109-8

Sloutsky, V. M. (2018). Category learning and conceptual development. In J. T. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (1st ed., pp. 1–46). Wiley. https://doi.org/10.1002/9781119170174.epcn402

Smith, E. E., Langston, C., & Nisbett, R. E. (1992). The case for rules in reasoning. *Cognitive Science*, *16*(1), 1–40. https://doi.org/10.1207/s15516709cog1601_1

Smith, E. E., & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition*, *22*(4), 377–386. https://doi.org/10.3758/BF03200864

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1–2), 159–216. https://doi.org/10.1016/0004-3702(90)90007-M

Song, S., Miller, K. D., & Abbott, L. F. (2000). Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience*, *3*(9), 919–926. https://doi.org/10.1038/78829

Spens, E., & Burgess, N. (2024). A generative model of memory construction and consolidation. *Nature Human Behaviour*, *8*(3), 526–543. https://doi.org/10.1038/s41562-023-01799-z

Starkweather, C. K., & Uchida, N. (2021). Dopamine signals as temporal difference errors: Recent advances. *Current Opinion in Neurobiology*, *67*, 95–105. https://doi.org/10.1016/j.conb.2020.08.014

Struyf, D., Zaman, J., Vervliet, B., & Van Diest, I. (2015). Perceptual discrimination in fear generalization: Mechanistic and clinical implications. *Neuroscience & Biobehavioral Reviews*, *59*, 201–207. https://doi.org/10.1016/j.neubiorev.2015.11.004

Sutton, R. S., & Barto, A. (2020). *Reinforcement learning: An introduction* (Second edition). The MIT Press.

Suzuki, M., Pennartz, C. M. A., & Aru, J. (2023). How deep is the brain? The shallow brain hypothesis. *Nature Reviews Neuroscience*, *24*(12), 778–791. https://doi.org/10.1038/s41583-023-00756-z

Tacikowski, P., Kalender, G., Ciliberti, D., & Fried, I. (2024). Human hippocampal and entorhinal neurons encode the temporal structure of experience. *Nature*, *635*(8037), 160–167. https://doi.org/10.1038/s41586-024-07973-1

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, *19*(1), 109–139. https://doi.org/10.1146/annurev.ne.19.030196.000545

Tang, E., Mattar, M. G., Giusti, C., Lydon-Staley, D. M., Thompson-Schill, S. L., & Bassett, D. S. (2019). Effective learning is accompanied by high-dimensional and efficient representations of neural activity. *Nature Neuroscience*, *22*(6), 1000–1009. https://doi.org/10.1038/s41593-019-0400-9

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. https://doi.org/10.1017/S0140525X01000061

Thornton, M. A., & Tamir, D. I. (2024). Neural representations of situations and mental states are composed of sums of representations of the actions they afford. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-024-44870-7

Thoroughman, K. A., & Shadmehr, R. (2000). Learning of action through adaptive combination of motor primitives. *Nature*, *407*(6805), 742–747. https://doi.org/10.1038/35037588

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522. https://doi.org/10.1038/381520a0

Thorpe, S. J. (2009). The speed of categorization in the human visual system. *Neuron*, *62*(2), 168–170. https://doi.org/10.1016/j.neuron.2009.04.012

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Tsunoda, K., Yamane, Y., Nishizaki, M., & Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience*, *4*(8), 832–838. https://doi.org/10.1038/90547

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience*, *21*(10), 1934–1945. https://doi.org/10.1162/jocn.2009.21131

Tversky, A., & Shafir, E. (Eds.). (2004). *Preference, belief, and similarity: Selected writings*. MIT Press.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, *5*(7), 682–687. https://doi.org/10.1038/nn870

Van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211–219. https://doi.org/10.1016/j.tins.2012.02.001

Vyas, S., Golub, M. D., Sussillo, D., & Shenoy, K. V. (2020). Computation through neural population dynamics. *Annual Review of Neuroscience*, *43*(1), 249–275. https://doi.org/10.1146/annurev-neuro-092619-094115

Wakhloo, A. J., Sussman, T. J., & Chung, S. (2023). Linear Classification of Neural Manifolds with Correlated Variability. *Physical Review Letters*, *131*(2), 027301. https://doi.org/10.1103/PhysRevLett.131.027301

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage*, *137*, 188–200. https://doi.org/10.1016/j.neuroimage.2015.12.012

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868. https://doi.org/10.1038/s41593-018-0147-8

Wang, J., Yang, Y., Fan, L., Xu, J., Li, C., Liu, Y., Fox, P. T., Eickhoff, S. B., Yu, C., & Jiang, T. (2015). Convergent functional architecture of the superior parietal lobule unraveled with multimodal neuroimaging approaches: Parcellation of superior parietal lobule. *Human Brain Mapping*, *36*(1), 238–257. https://doi.org/10.1002/hbm.22626

Wang, W., & Carreira-Perpinan, M. (2014). The role of dimensionality reduction in classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *28*(1). https://doi.org/10.1609/aaai.v28i1.8975

Warden, M. R., & Miller, E. K. (2010). Task-dependent changes in short-term memory in the prefrontal cortex. *Journal of*

Neuroscience, *30*(47), 15801–15810. https://doi.org/10.1523/JNEUROSCI.1569-10.2010
Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3–4), 279–292. https://doi.org/10.1007/BF00992698
Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541. https://doi.org/10.1038/s41562-023-01659-w
Whitlock, J. R. (2017). Posterior parietal cortex. *Current Biology*, *27*(14), R691–R695. https://doi.org/10.1016/j.cub.2017.06.007
Whittington, J. C. R., Dorrell, W., Ganguli, S., & Behrens, T. E. J. (2022). *Disentanglement with Biological Constraints: A Theory of Functional Cell Types*. https://arxiv.org/abs/2210.01768v2
Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. J. (2020). The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell*, *183*(5), 1249-1263.e23. https://doi.org/10.1016/j.cell.2020.10.024
Wikenheiser, A. M., & Schoenbaum, G. (2016). Over the river, through the woods: Cognitive maps in the hippocampus and orbitofrontal cortex. *Nature Reviews Neuroscience*, *17*(8), 513–523. https://doi.org/10.1038/nrn.2016.56
Wimmer, G. E., Daw, N. D., & Shohamy, D. (2012). Generalization of value in reinforcement learning by humans. *European Journal of Neuroscience*, *35*(7), 1092–1104. https://doi.org/10.1111/j.1460-9568.2012.08017.x
Wiskott, L., & Sejnowski, T. J. (2002). Slow Feature Analysis: Unsupervised Learning of Invariances. *Neural Computation*, *14*(4), 715–770. https://doi.org/10.1162/089976602317318938
Wolpert, D. M., Diedrichsen, J., & Flanagan, J. R. (2011). Principles of sensorimotor learning. *Nature Reviews Neuroscience*, *12*(12), 739–751. https://doi.org/10.1038/nrn3112
Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, *114*(2), 245–272. https://doi.org/10.1037/0033-295X.114.2.245
Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244
Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111
Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, *22*(2), 297–306. https://doi.org/10.1038/s41593-018-0310-2
Zaki, S. R., & Salmi, I. L. (2019). Sequence as context in category learning: An eyetracking study. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, *45*(11), 1942–1954. https://doi.org/10.1037/xlm0000693

Zaman, J., Chalkia, A., Zenses, A.-K., Bilgin, A. S., Beckers, T., Vervliet, B., & Boddez, Y. (2021). Perceptual variability: Implications for learning and generalization. *Psychonomic Bulletin & Review*, *28*(1), 1–19. https://doi.org/10.3758/s13423-020-01780-1

Zhong, L., Baptista, S., Gattoni, R., Arnold, J., Flickinger, D., Stringer, C., & Pachitariu, M. (2024). *Distinct streams for supervised and unsupervised learning in the visual cortex*. https://doi.org/10.1101/2024.02.25.581990

Zhou, J., Jia, C., Montesinos-Cartagena, M., Gardner, M. P. H., Zong, W., & Schoenbaum, G. (2021). Evolving schema representations in orbitofrontal ensembles during learning. *Nature*, *590*(7847), 606–611. https://doi.org/10.1038/s41586-020-03061-2