Exploring Generalization Inductive Biases in the Brain and Deep Neural Networks:

Experimental and Computational Approaches

By

MARYAM ZOLFAGHAR

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Randall C. O'Reilly, Chair

Hamed Pirsiavash

Erie Boorman

Committee in Charge

2023

ABSTRACT

Exploring Generalization Inductive Biases in the Brain and Deep Neural Networks: Experimental and Computational Approaches

Maryam Zolfaghar

What enables humans to effortlessly learn and generalize across diverse tasks, an exceptional ability that even advanced deep neural networks struggle to replicate? This thesis explores this question, crucial in both neuroscience and computer science. Despite deep neural networks' achievements, they require vast data to learn, sometimes even surpassing human lifetimes' worth of experience. In contrast, humans adapt existing knowledge to novel challenges, indicating the presence of cognitive mechanisms facilitating adaptive out-of-domain generalization. This thesis comprehensively explores some of these mechanisms, including how humans learn in a way that supports the development of representations applicable across diverse contexts, adjust and deploy them based on contextual demands, and form such representations over different time scales, especially when rapid learning is required, and how it achieves such learning without forgetting the past knowledge.

The human brain has a remarkable ability to predict what will happen next and subsequently compares these predictions to what actually happens. The differences form prediction errors. These errors serve as self-generated teaching signals that guide us in adjusting our understanding and mental models to better match reality. This process, known as deep predictive learning, helps us adapt and refine our internal representations, fostering adaptable abstract representations that capture patterns within sensory inputs, which are central for out-of-domain generalization. Additionally, our brain forms abstract map-like representations (i.e., cognitive maps) that empower advanced reasoning skills and bridge learned knowledge with novel challenges, facilitating successful problem-solving. Cognitive control mechanism within the prefrontal cortex, known for systematic generalization, dynamically interacts with these representations to meet multiple objectives. This capability is instru-

mental in navigating complex tasks, generalizing to unfamiliar situations, and embracing new challenges. Constructing such abstract representations demands gradual experience integration. However, according to the CLS framework, the brain also possesses mechanisms for when rapid learning is required in new environments without catastrophically forgetting the prior knowledge.

Generalization is a cornerstone of human intelligence, enabling us to tackle daily and novel challenges. This thesis enriches our understanding through a multidisciplinary approach and highlights the integration of experimental and computational techniques. It combines neural EEG and behavioral data with machine learning models to explore the predictive learning process. Additionally, it builds deep neural networks to replicate fMRI findings, with a particular emphasis on cognitive control processes associated with the generation of map-like representations. The study quantifies generalization in these networks, introduces cognitive-inspired inductive biases to these models, and develops models consistent with the CLS framework for tasks requiring generalization across various time scales. By merging computational and experimental methods, this research offers insights into scenarios challenging to replicate with human participants, in addition to inspiring the development of advanced models and contributing to the ongoing evolution of future AI systems.

Copyright

Acknowledgements

Please note that specific acknowledgments for each manuscript can be found at the end of their respective chapters.

Individuals and Groups

I would like to thank my advisor, Randall C. O'Reilly, for giving me the invaluable opportunity to be a part of his research lab. His guidance and support have been essential throughout my Ph.D. Without his mentorship and insights, I wouldn't have been able to achieve what I have.

I am also thankful to the other members of my committee, Erie Boorman and Hamed Pirsiavash, for their valuable comments and feedback. I would also like to thank my colleagues at the Computational Cognitive Neuroscience Lab: Jake Russin, John Rohrlich, Thomas Hazy, Kai Kruger, Ananta Nair, Seth Herd, David Noelle, April Luo, Kevin McKee, Will Chapman, Andrew Carlson, Riley DeHaan, Jessica Mollick, Dean Wyatte, and Wolfgang M. Pauli. Our discussions and collaborations over the years have been invaluable. I extend a special thanks to Steve Luck for his invaluable advice and insightful comments, which have greatly influenced and shaped my ideas. I also want to express my appreciation to the members of the Learning and Decision Making Lab, including Seongmin (Alex) Park and Sarah Sweigart, as well as the members of the Luck Lab, including Aaron Matthew Simmons.

I am also grateful to Tim Curren and Lewis O. Harvey, Jr. at CU Boulder. Their profound expertise, particularly in the field of experimental research, along with their valuable

guidance and insightful suggestions, have consistently enhanced the quality of my research. Additionally, I want to express my thanks to the members of the Curran Lab, including Bill Carpenter and Levi Davis.

My appreciation extends to the members of the Analogy Group, namely Jonathan Cohen, Alexander Petrov, Tim Buschman, Taylor Webb, Steven Frankland, Sebastian Musslick, Zachary Dulberg, Tyler Giallanza, Simon Segert, and Randy Gobbel – for their insightful discussions and perspectives. Additionally, I am grateful to the members of the Memory Meeting at UC Davis, particularly Charan Ranganath, and to the members of Serre Lab at Brown University, including Thomas Serre, Drew Linsley, Lore Goetschalckx, Lakshmi N. Govindarajan, Alekh Karkada Ashok, and Rex Liu, for their valuable insights and collaborations. A special thank you also goes to my friends Maya, Hoss, and Aliz for their continuous support and presence over the years.

Overall, I am grateful for the meaningful discussions, guidance, and collaborations that I have had with all of these individuals. Their contributions have significantly contributed to the development of my research and enriched my graduate school experience.

Last but certainly not least, I would like to express my heartfelt gratitude to my parents, Hamideh and Mohsen, and my sisters, Marjan and MahYas, for their unwavering support throughout these years. Their constant presence and encouragement have been my pillars of strength, guiding me every step of the way. Without them, I would not have been able to move forward on this path, and there are no words to truly capture my deep appreciation for them. I am thankful for all that they have done for me.

Funding

The work in this dissertation was supported in part by: ONR grants ONR N00014-20-1-2578, N00014-19-1-2684 / N00014-18-1-2116, N00014-18-C-2067, ONR N00014-14-1-0670 / N00014-16-1-2128, as well as NSF CAREER Award 1846578, and NIH R56 MH119116.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Table of contents

List of Tables

List of Figures

Dedication

To my parents, Mohsen and Hamideh. Your steadfast support and guidance have made you my greatest mentors in life. Thank you for always believing in me.

Chapter 1

Introduction

What are the underlying mechanisms that result in the ability to generalize out of domain? Answering this central question has been of great importance in many lines of research in neuroscience and computer science. Although deep neural networks have recently received tremendous success on challenging machine learning tasks, these models can only achieve high performance when trained on a vast amount of training data identically and independently distributed (i.i.d.). However, these models struggle to solve novel problems that are drawn from a different distribution (*out-of-distribution*, o.o.d generalization) (Lake et al., 2017), which also lines up with the data inefficiency of large neural networks. These networks can encounter much more information than humans experience in a lifetime. For instance, GPT-2's training is based on data (∼8 billion tokens of internet text (Radford et al., 2019)) that is comparable to the content of 10 human lifetimes. In comparison, the training of its successor, GPT-3, utilizes a dataset (∼114 billion words broken into ∼500 billion subword tokens (Brown et al., 2020)) equivalent to the knowledge from 100 human lifetimes (Wilcox et al., 2022). Unlike these models, humans can generalize to perform novel tasks that they have never encountered before, even if they are only given a few external teaching signals, suggesting that the brain contains mechanisms that support such generalization.

In summary, this research makes a contribution to our understanding of the mechanisms underlying predictive learning, a fundamental learning process underlying the development of high-level abstract representations that support generalization. Additionally, our findings

highlight the role of specific biological properties of the neocortex in generating abstract and shared representations, supporting the formation of the structural basis for systematic generalization. This process is closely intertwined with cognitive control mechanisms that support the dynamical and flexible modulation of the shared neocortical representation according to the objective of the given task, which is essential in resolving potential conflicts and preventing catastrophic interference. Furthermore, in situations when rapid learning in a new environment is necessary, another brain region, the hippocampus, becomes involved. According to the CLS framework, rapid learning within the hippocampus is attributed to its pattern-separated and sparse representations. This mechanism complements the neocortical processes and offers insights into how the brain efficiently and rapidly adapts to new environments.

In our study, we employed a comprehensive approach involving experiments, EEG neural signals, and machine learning models to delve deeply into the brain patterns associated with predictive learning. Furthermore, we integrated human functional magnetic resonance imaging (fMRI) with deep neural networks to explore abstract representations that exhibit map-like characteristics, particularly in tasks necessitating generalization beyond familiar domains. Our research also delved into the role of cognitive control in leveraging such map-like abstractions to enhance generalization and address interferences. Expanding upon these findings, we then introduced a deep neural network incorporating explicit cognitive control processes. This network's capacity for continual learning was assessed in both blocked and interleaved learning schemes. Finally, we delved into two complementary systems—slow learning and fast learning—that synergistically promote generalization through the neocortex and hippocampus. We developed consistent deep neural networks guided by insights derived from the complementary learning systems (CLS) framework. The subsequent sections delve into more details of each of these mechanisms.

In the brain, the ability to generalize appears to rely on constructing high-level abstract representations of the environment that can later be used while solving novel tasks (Park

et al., 2020; Whittington et al., 2020; Behrens et al., 2018). Recent studies in neuroscience explored the mechanisms by which the brain forms abstract representations that can capture the underlying structure of the environment (i.e., cognitive maps), and this map-like representation is thought to be utilized to generalize to a novel sample (Park et al., 2020; Whittington et al., 2020; Behrens et al., 2018). Like geographic maps showing the actual distance between places (accounting for scale), cognitive maps capture the latent structure and relations of the task, and therefore, distances in the latent representational space of these maps reflect the distances in the feature space.

The development of such representations occurs slowly over a long time and requires a wide range of experience across various tasks (McClelland et al., 1995; O'Reilly et al., 2011). However, the formation of such representations might not require a massive amount of external teaching signals to train the brain at any given time in a supervised manner.

More recent research (O'Reilly et al., 2021b, 2014, 2013; Wyatte, 2014) suggested that the brain develops high-level abstract representations while learning from nothing but predicting the subsequent inputs (i.e., *predictive learning*). Specifically, the brain is constantly predicting the following sensory input, and the discrepancies between its predictions and the actual input (referred to as *prediction error*) drive learning, which improves predictions over time (O'Reilly et al., 2021b; Bullier, 2001; Friston, 2005). Therefore, through internal predictions, the brain teaches itself even in the absence of external teaching signals. Based on the different phases involved in predictive learning, which encompass prediction and subsequent outcome, there must be a mechanism by which the brain switches between these two states, which is represented over separate time scales. One prominent theory (O'Reilly et al., 2021b, 2017) suggests that this separation occurs temporally in 100 ms intervals: the brain generates its prediction during the first 75 ms, and then in the last 25 ms, when it experiences the actual sensory input, the brain updates its internal representation as a function of the prediction errors. It has been shown that through this process of predicting the following sensory inputs and without any additional external teaching information, abstract and

high-level representations emerge from an ongoing stream of perceptual experience (O'Reilly et al., 2017, 2021b).

Recent studies suggest that the brain constructs cognitive maps both in spatial tasks (e.g., navigation) and non-spatial reasoning tasks where people have to reason about novel relationships between the inputs (Behrens et al., 2018; O'Reilly et al., 2021a; Summerfield et al., 2020; Whittington et al., 2020). Most of these studies include tasks in which the decision must be made within a single context or goal. However, humans are often involved in tasks that require decisions to be made by selecting relevant stimulus features that satisfy multiple possible objectives. These scenarios may introduce conflict or interference that requires a mechanism to satisfy one objective while resolving conflicts from other objectives. Despite a recent impressive performance in deep neural networks, these models perform poorly when multiple goals should be met at each moment, and they overwrite the information of the previously learned task (i.e., *catastrophic forgetting*). In other words, these models catastrophically forget the knowledge from the prior tasks while learning to perform in a novel environment.

The brain's capacity to dynamically and flexibly select and attend to task-relevant features to avoid interference and facilitate learning a new task without catastrophic forgetting is known as cognitive control (Miller & Cohen, 2001; Herd et al., 2006; Rougier et al., 2005). Classical theoretical work (Miller & Cohen, 2001) suggests that cognitive control is implemented as a top-down attentional signal that modulates information in other brain areas. More specifically, it has been proposed that the prefrontal cortex dynamically modulates processing in more posterior regions. These signals suppress task-irrelevant processes and excite task-relevant ones (Aben et al., 2020). However, less is known about how cognitive control might influence map-like representations, such as those observed in the medial temporal lobe (MTL) or parietal cortex during abstract relational reasoning tasks.

Previous work illustrated that the mechanisms involved in generalization depend on the nature of the task environment (O'Reilly et al., 2011). As we mentioned earlier, the develop-

ment of map-like representations occurs slowly and integrates across many experiences and, therefore, can be leveraged to make novel inferences. For example, in a familiar and well-learned environment like one's hometown, it is easy to navigate to a new place by reasoning over past experiences. Also, when visiting a new city, it is still possible to navigate to a novel location by reasoning over recent experiences, even those that integrated on the same day. Therefore, there might be two learning systems in the brain underlying relational reasoning and generalization abilities.

The well-supported complementary learning systems (CLS) framework introduces two qualitatively different neural mechanisms underlying generalization abilities in familiar and novel environments (McClelland et al., 1995; O'Reilly et al., 2011). The CLS framework suggests two learning mechanisms unfolding over two different timescales, which involve separate brain regions. This framework explains how the brain can support learning representations without catastrophically forgetting previous ones. In this framework, the neocortex learns slowly, creating more abstract representations that accumulate over many experiences and can be utilized to make novel decisions. However, this type of learning may not be possible in a new environment when there are not enough experiences and where samples are not interleaved to allow for efficient adaptation. From the CLS perspective, in a novel environment, fast learning occurs in the hippocampus due to its sparse and pattern-separated representations. Having little to no overlap across the representations allows the hippocampus to learn novel samples without catastrophic interference.

In summary, developing more abstract and shared representations in the neocortex, which builds up structural representations, provides the foundation for systematic generalization. Furthermore, cognitive control can dynamically and flexibly modulate this representation to meet the demands of the current goal (Miller & Cohen, 2001; Herd et al., 2006; Rougier et al., 2005) and in a way that solves conflicts and avoids catastrophic interference. Additionally, another complementary mechanism that supports generalization in a more unfamiliar environment where rapid learning is required is storing sparse representations in the

hippocampus, which has also been shown previously to play a role in rapid generalization (Eichenbaum, 2004; Zeithamova et al., 2012).

The work presented in the following chapters utilizes experimental and computational approaches to explore the brain's underlying mechanisms and inductive biases that support systemic out-of-domain generalization. Specifically, chapter 2 explores the mechanisms of predictive learning by using a combination of behavioral experiments, neural EEG signals, and machine learning models. Subsequently, chapter 3 and chapter 4 combine human functional magnetic resonance imaging (fMRI) with deep neural network models to study the formation of map-like representations and cognitive control mechanisms. Within chapter 3, we investigate the interplay between cognitive control and the geometry of map-like representations. Building upon this foundation, chapter 4 introduces a neural network model equipped with an explicit control mechanism, providing insights into empirical findings of continual learning over blocks vs. interleaved learning. Finally, chapter 5 investigates the synergy between the gradual neocortical and rapid hippocampal generalization with insights from the complementary learning system framework. In the following sections, we will go over each of these chapters with more details.

First, chapter 2 (Zolfaghar et al., In-prep) explores the learning mechanism (i.e., predictive learning) supporting the formation of high-level abstract representation and subsequently systematic generalization. In this chapter, two behavioral and neurophysiological experiments and their results are presented to assess the underlying mechanisms associated with predictive learning. In this work, I carefully designed the experiments to isolate pure perceptual learning (Coomans et al., 2012; Foerde & Poldrack, 2016; Clegg et al., 1998; Willingham, 1999; Willingham et al., 1989; Dennis et al., 2006; Deroost & Soetens, 2006) and facilitate implicit learning (Cleeremans & McClelland, 1991; Berry & Dienes, 1993; Shanks, 2005; Daltrozzo & Conway, 2014) among participants using statistical-sequential learning paradigms (Summerfield & de Lange, 2014). The brain EEG signals were recorded in collaboration with Steve Luck's lab at UC Davis.

I employed multifaceted approaches and combined both experimental and computational tools to explore the proposed research question. While different techniques were available, I opted for more straightforward methods that offered the most relevant insights into the data. One such technique is decoding, which can directly utilize neural information even from the brain's surface (Bae & Luck, 2018; Luck, 2022). Even though decoding has been used using various types of neural data (Serences et al., 2009; Harrison & Tong, 2009; Ester et al., 2013; Wolff et al., 2017; Bae & Luck, 2018; Rihs et al., 2007; LaRocque et al., 2013; Rose et al., 2016; Foster et al., 2017; van Ede et al., 2017; Bae & Luck, 2018), to preserve optimal temporal resolution, we chose to use ERP signals for our decoding analysis (de Cheveigné & Nelken, 2019). Furthermore, alongside decoding, we applied a more generalized approach known as temporal generalization (King & Dehaene, 2014; King et al., 2014). It is an extension of the decoding over time approach and generates cross-temporal matrices as a result that shows the evolution of decoding patterns over time and can better make a connection between the information of the pre-and post-stimulus time windows. The representation of the decoded information is further explored by using representational similarity analysis (RSA) (Nili et al., 2014). RSA is a method used in neuroscience to compare and analyze the patterns of brain activity evoked by different experimental conditions or trial types. In our study, we employed RSA to determine whether the representation of a trial is more similar to the representation of the current location on the screen, or if it aligns more closely with the representation of the previous or next locations.

Consistent in both experiments, participants showed pure perceptual implicit learning, both a general speed-up in task acquisition and a specific sequence learning. Decoding analyses consistently demonstrated high decoding accuracy for future locations throughout the duration of our epoch. Employing a generalized decoding approach that formed a cross-temporal matrix, further supported this finding and made a connection between the past and future time points. We observed that our decoder, trained on the pre-stimulus time window, can accurately predict the next locations when tested on the post-stimulus time points. These

results collectively suggest the utilization of the predictive information present in our task even though the participants were not explicitly informed about such predictive relationships. In our RSA approach, this was further supported by a sustained similarity activity with the current location that persisted until the presentation of the next trial. This persistence was essential for successfully predicting the next locations, due to the second-order characteristic of our sequences. This characteristic requires the retention of the second-order information (from two previous locations) to effectively predict the upcoming location. This observation further validates the effective utilization of predictive relationships within our second-order sequences to anticipate future locations.

Subsequently, in chapters 3 and 4 (Zolfaghar et al., 2022; Park A. et al., In-prep; Russin et al., 2022), a combination of a human fMRI study with a deep neural network presented to assess the constructed high-level representations (i.e., cognitive maps). In this context, both humans and the neural network model learned the same task, which required making transitive inferences. This work has been done in collaboration with Erie Boorman's fMRI lab at UC Davis. The findings showed that even though the model and participants were not explicitly told about the latent 4×4 structure of the task, representations in the hidden layers of the model and brain regions, including the hippocampus (HC), entorhinal cortex (EC), and orbitofrontal cortex (OFC) captured this basic structure. Furthermore, the relationship between cognitive control and map-like representations is explored. The task used here was designed to facilitate the learning of map-like representations while simultaneously requiring the use of cognitive control as a function of the current task context. Our results showed that in addition to the 2D map-like representations found in our previous results, these representations' geometry was warped along the congruent context due to the demand for cognitive control when there was interference. In other words, the pattern similarity between incongruent pairs was greater than that of congruent pairs. Moreover, we observed such warped representations in the hidden representations of the model and in HC, medial prefrontal cortex (mPFC), and amygdala. Furthermore, these representations were dynam-

ically modulated by cognitive control to resolve the conflict caused by incongruence in the task, which might be another way to avoid interference. In so doing, the irrelevant context of the representational space was suppressed according to the current context. We found such 1D representations in the model and in brain regions, including the posteromedial cortex (PMC) and dorsomedial frontal cortex (dmFC).

Chapter 3 (Zolfaghar et al., 2022) studies the connections between control processes and the geometry of cognitive maps learned by neural networks. Building upon this work, chapter 4 (Russin et al., 2022) introduces a neural network model that incorporates an explicit control mechanism. This model effectively explains empirical findings related to blocked versus interleaved learning (Flesch et al., 2018).

Developing map-like representations requires integrating information across many experiences over a long period. For example, in the fMRI study, participants were trained on the task for three consecutive days to be able to perform well. However, this slow learning is not possible in environments where learning occurs too quickly, or sequences of events are not sufficiently interleaved. For example, participants in the fMRI study performed the last part of the training on the third day and also had to make transitive inferences on the same day. Therefore, the brain might have multiple mechanisms that support generalization in different situations. Additionally, a repetition suppression analysis found a retrieval of particular faces (i.e., hubs) in the hippocampus. These hubs were necessary to find the correct answer on the third day at the time of inference. These results suggest that complementary to the map-like representations found in the neocortex, the hippocampus is involved when learning occurs too quickly.

Finally, in chapter 5 (Russin et al., 2021), we investigate the interplay between slow generalization in the neocortex and rapid generalization in the hippocampus by applying the principles of the CLS framework. We then built deep neural networks to capture the basic properties of this framework. Finally, we evaluated our models of the cortical and episodic memory systems using the transitive inference task introduced in the previous chapters,

which involved learning over two different time scales. Participants were trained over multiple days, and also they were trained on samples given on the same day as the transitive inference test. We found that both models are capable of solving the transitive inference task and reproduced key findings from the fMRI experiment (Park et al., 2020).

Chapter 2

From Past to Future: Exploring predictive learning in the Brain Using Computational and Experimental Approaches[1]

Maryam Zolfaghar[2] [3], Chaodan Luo[3], Aaron Simmons[4], Tim Curren[5], Steven J. Luck[4], Randall C. O'Reilly[3]

2.1 Abstract

How does the brain create complex abstract knowledge from continuous sensory information, even without explicit teaching cues? The brain has the remarkable capacity to learn passively by observing continuous inputs and actively generating predictions about upcoming information. It then refines its internal models by minimizing the discrepancies between these predictions and real-world experiences (i.e., prediction errors), referred to as predictive learning. This mechanism enables the brain to use its prediction errors as an internal source of teaching signals. In this study, we hypothesize that the two phases involved in predictive learning (prediction and actual outcomes) consist of a rapid sequence of prediction-outcome cycles that unfold across distinct temporal windows. The brain first generates predictions, and after encountering the actual outcome, it refines its predictions to reduce its errors. To explore this learning process, our study employs a combination of behavioral experiments and electroencephalography recordings. Participants exhibit learning not only in terms of

[1*] A version of this article is going to be submitted for publication.

[2] Department of Computer Science, University of California, Davis
[3] Center for Neuroscience, University of California, Davis
[4] Center for Mind and Brain, University of California, Davis
[5] Department of Psychology and Neuroscience, University of Colorado, Boulder

general task enhancement but also in the context of sequence-based learning. Notably, this learning is purely perceptual and occurs without participants' awareness, implying implicit learning. EEG decoding methods revealed the presence of significant decodable information about future locations in our sequences. A cross-temporal analysis further supports the existence of predictive learning by demonstrating significant decoding accuracy when a machine learning model is trained on pre-stimulus time points and tested on post-stimulus ones. Additionally, representational similarity analysis demonstrated the retention of the necessary information for predicting the next location until the presentation of that location—a further hallmark of predictive learning.

2.2 Introduction

Understanding how the human brain can transform continuous streams of sensory information into complex abstract knowledge without relying on explicit external teaching signals remains a fundamental question. This inquiry has motivated research that aims to understand how the brain can acquire such knowledge naturally and passively, solely through its interaction with sensory stimuli. An emerging hypothesis suggests the existence of an underlying mechanism within this passive learning process: an active network that generates predictions about upcoming sensory inputs, a learning process known as predictive learning (O'Reilly et al., 2021b). These predictions often deviate from reality, leading to what researchers term "prediction errors". It is these errors that drive the learning process (Bullier, 2001; Friston, 2005; O'Reilly et al., 2021b). Over time, as the brain works to minimize these prediction errors through an iterative process, it continually refines its internal models and representations of the surrounding environment.

This predictive learning mechanism not only enables the brain to comprehend its surroundings but also plays a crucial role in the formation of abstract representations. These representations are characterized by their ability to capture the underlying structure and

regularities of sensory inputs. As a consequence, they provide a foundation that equips the brain to effectively adapt to novel and unseen situations (O'Reilly et al., 2021b, 2017).

Unlike earlier deep neural networks, which often require numerous labeled training samples to address specific tasks and still struggle to generalize that learning to new tasks (Lake & Baroni, 2018; Lake et al., 2019), there is a noticeable shift towards employing learning mechanisms more consistent with predictive learning in these models (Devlin et al., 2019; Radford et al., 2019; Peters et al., 2018). This entails adopting approaches that rely less on external teaching signals. Large language models (LLMs) exemplify this shift, being trained through predictive learning rather than conventional reliance on labeled data (Radford et al., 2019; Brown et al., 2020; Wilcox et al., 2022). This demonstrates how the incorporation of predictive capabilities can transform the potential of advanced deep neural networks, which already have been evolved in the brain, albeit with some differences. While LLMs are trained with an emphasis on parallel processing, the brain engages in sequential learning through prediction-outcome sequences that unfold across different temporal windows (O'Reilly et al., 2021b). Despite disparities, the core principle is shared — these models fundamentally predict what is coming next, aligning them computationally with predictive learning. This brings forth the question: How does this influential form of predictive learning operate within the brain?

Predictive learning has emerged as a robust framework for deciphering how the brain learns from its environment and processes sensory information (Friston, 2005; Mumford, 1992; Rao, 1999; O'Reilly et al., 2014) through intrinsic and ongoing predictions and its continuous effort to minimize the mismatches between these predictions and real-world experiences. This framework implies that predictive learning could serve as a potent source of essential learning signals for human intelligence (Bullier, 2001; Friston, 2005; Miall & Wolpert, 1996). However, less is known about how predictive learning works in the brain.

Since this learning dynamic is based on different phases (predictions and actual outcome), there should be a mechanism in the brain to switch between these two states (Friston, 2005;

Bullier, 2001; O'Reilly et al., 2021b, 2017). One prominent theory suggests that these two states occur separately in bottom-up and top-down connections of the brain (Friston, 2005). In this hypothesis, the prediction error is explicitly constructed by comparing the prediction (conveyed by top-down connections) with actual neural activities (transmitted by bottom-up pathway). Contrary to what Friston (2005) suggested, we have hypothesized that prediction error is implicitly represented as a *temporal difference* (O'Reilly et al., 2021b, 2014). First, the brain generates its predictions. Then later, when the brain experiences the actual sensory stimulus, the temporal difference between the predictions versus the ground-truth sensory signal (i.e., *prediction error*) updates the internal representation to reduce the prediction error. This theory suggests that, based on the biological evidence (Andersen & Andersson, 1968; Hughes et al., 2004), the brain's predictions are generated at 100 ms (i.e., alpha frequency, 10 Hz) intervals (O'Reilly et al., 2021b, 2014, 2013; Wyatte, 2014). Specifically, within the first 75 ms of the overall 100 ms, the brain generates its predictions. Then, when the actual sensory inputs are experienced over the last 25 ms of the 100 ms, any discrepancies between this ground-truth outcome and the previous prediction state are the teaching signals that propagate throughout the brain to shape its representations.

Our study puts the alternative hypothesis to the test through two experiments where participants implicitly and perceptually learn to predict upcoming stimuli. We used a combination of the pure perceptual learning (Coomans et al., 2012) and statistical-sequential learning (Summerfield & de Lange, 2014) paradigms. In these paradigms, unbeknown to participants, there is an underlying statistical structure to the task. Therefore, throughout learning, participants may implicitly learn the hidden structural rule leading to faster and more accurate responses. To measure the extent to which such learning is due to learning the underlying sequence rather than a general learning trend in our behavioral study, we replaced the underlying pattern with another sequence after participants learned to perform the task well. Therefore, if participants learn the underlying pattern, replacing the sequence with another one should increase their reaction times in response to the stimulus. Alterna-

tively, with no such sequential learning, we may expect to see a decrease in reaction times throughout the whole experiment, solely due to an enhancement in familiarity with the task and an improved understanding of the overall paradigm.

Our findings revealed that participants learned to respond faster and more accurately throughout the test session in both experiments. In addition to this generic form of learning, we also discovered evidence that this learning is attributed to acquiring knowledge of the underlying sequence. This was apparent from the observed increase in participants' reaction times when the underlying sequential pattern was violated. More specifically, we observed that participants' reaction times increased when we transitioned from a more frequent sequence (i.e., primary sequence) to a less frequent one (i.e., alternative sequence). Conversely, when we reverted to the primary sequence, their reaction times decreased once again. Statistical analysis confirmed the significance of this observed pattern.

Moreover, to measure the extent to which this sequence-based learning occurred implicitly, we included supplementary tasks at the end of both experiments. The results from these tasks provided further evidence that this learning occurred without participants being consciously aware of the underlying sequences, even though their brains implicitly exhibited signs of sequence learning.

The consistency in behavioral findings across both experiments underscores that participants implicitly gained knowledge of the underlying statistical sequence solely through pure perceptual learning. Therefore, to more precisely examine the brain signatures underlying predictive learning, we employed electroencephalography (EEG) recordings of brain signals in the second experiment. EEG is particularly suitable for tracking time-sensitive changes in the brain. Our primary objective was to uncover the specific brain patterns associated with its capacity to predict future occurrences based on preceding events. Initially, we employed decoding analysis with the main goal of exploring the patterns linked to the brain's ability to predict subsequent locations.

To accomplish this, we organized our trials into four distinct groups according to the trial

location at time = 0 ms. Within each group, the preceding and subsequent trials had two potential locations each (i.e., two distinct bins within each group), due to the second-order feature of our sequences (Coomans et al., 2012). This grouping strategy makes the chance of predicting the next location 50% within each bin. Throughout our analysis, we referred to the trials presented immediately before time = 0 ms as 'previous trials,' those after as 'next trials,' and the trial at time = 0 ms as the 'current trial.'

We used standard second-order sequences that are widely-used in the sequence-learning literature (42132431 and 13423124), which make it impossible to predict the next stimulus in the sequence based only on the current stimulus. Instead, two prior stimuli ("second-order information") are required to accurately predict the next stimulus. While this is critical for making it more difficult to learn the sequences, and ensure that something beyond simple pairwise associative learning is required, it also means that it is more difficult to distinguish a representation of the previous stimulus from that of the next stimulus, because these are always linked (despite being different locations). For example, for a current trial of location 1, the second-order sequence contains either `'2 -> 1 -> 3'` or `'3 -> 1 -> 4'`, such that the 2 and 3 as previous and next are linked, as are 3 and 4. Thus, while our analyses identify a clear ability to predict the next stimulus in a sequence, it is not possible to distinguish this from the ability to remember the previous item, which is logically required for prediction in the first place.

Our findings revealed a significant decoding accuracy that persisted throughout the epoch time, showing that the brain had significant decodable information regarding the previous and next locations. To further examine these decoding results, we used a cross-temporal matrix that tested whether the decoding model is capable of using past information to predict future events. In other words, if we train the model on the information from the past, the model should be able to decode the information from the future. Indeed, we found that our model was able to decode the next stimuli when trained during the pre-stimulus time window and tested during the post-stimulus period, consistent with the predictive learning

hypothesis.

Building on the encouraging decoding and cross-temporal findings, which revealed the presence of decodable information associated with predictive learning, we delved deeper into the nature of this decoded information using representational similarity analysis (RSA) that examines the similarity structure of the representations over time (Nili et al., 2014). Interestingly, our RSA results showed a gradual shift in similarity, transitioning from previous-related information to current-related information, and then to next-related information as time progressed. This temporal progression of similarity closely aligned with the timing of our task. Furthermore, the RSA results also revealed a sustained similarity representation of the current trial, which is consistent with the demand for resolving the ambiguity present in the second-order sequences that we used, which requires the retention of information from the two previous locations for the successful prediction of the next location.

In the upcoming sections, we outline the methods employed in our study, provide explanations of the experimental paradigms, and delve into the behavioral findings. Next, we present the EEG results, beginning with our decoding analysis, which are further supported by the outcomes of the cross-temporal analysis, and then the RSA analyses. Finally, we will summarize our findings, discuss their relationship with other experiments and theoretical frameworks, and propose future work to address the limitations of the present study. Our overall goal is to contribute to a deeper understanding of the neural mechanisms driving predictive learning—a crucial cognitive process that facilitates the creation of adaptable internal models of our world and enables the capacity to generalize beyond specific domains.

2.3 Material and Methods

2.3.0.1 Participants

Experiment 1. A total of 31 volunteers took part in our first experiment (19 female, 12 male, mean age = 19). All participants were affiliated with the University of Colorado at

Boulder, and they either got credits for their introductory psychology course or received payment for their participation. They all were right-handed and reported to have normal or corrected-to-normal vision. Additionally, they were devoid of any psychiatric or neurological disorders and provided informed consent in accordance with the human subjects policy at CU Boulder. The Institutional Review Board granted approval for the study.

Experiment 2. A total of 40 volunteers took part in the experiment (33 female, 6 male, 1 non-binary, mean age = 20.48). One subject was removed from the analysis due to technical issues during EEG recording. Another subject was removed due to not having enough good trials after our artifact rejection. In addition, data from three subjects were collected during the pilot phase of the study that were not used for the analysis. All participants were from the University of California, Davis, and they all got paid for their participation. All participants were right-handed and indicated having either normal vision or vision corrected to a normal state. Moreover, every individual enrolled in the study was devoid of any psychiatric or neurological disorders and provided informed consent prior to the experiment in alignment with the human subjects policy at UC Davis. The study was approved by the Institutional Review Board.

2.3.0.2 Stimuli and apparatus

Participants executed all parts of the experiments in a semi-darkened cubicle of the psychological laboratory of the Colorado Boulder University for the first experiment and of the Center for Mind and Brain at UC Davis for the second experiment. Stimuli were paired letters, surrounded a fixation point. These pairs were either a target ("XO" or "OX") or a distractor ("YQ" or "QY"). Each trial was composed of one target and three distractors. The identities of targets and distractors always varied randomly with an equal probability. In the initial experiment, stimuli were arranged in a diamond shape, positioned at the clock locations of 3, 6, 9, and 12 o'clock (Figure 2.1A). A 17-inch LCD monitor displayed images

with a spatial resolution of 1280 × 1024 pixels and a refresh rate of 60 Hz. For the subsequent experiment, the letters were arranged in a square pattern, and we used a 24-inch LCD monitor (HP ZR2440W) with a spatial resolution of 1920 × 1200 pixels and a refresh rate of 60 Hz to present the trials. Under this configuration, in both experiments, the stimuli centered approximately .63° of visual angle from the center of the screen, and a chin rest was installed to maintain a fixed viewing distance and head position. In both experiments, stimuli were generated with MATLAB and presented with PsychToolbox package (Pelli, 1997; Brainard, 1997).

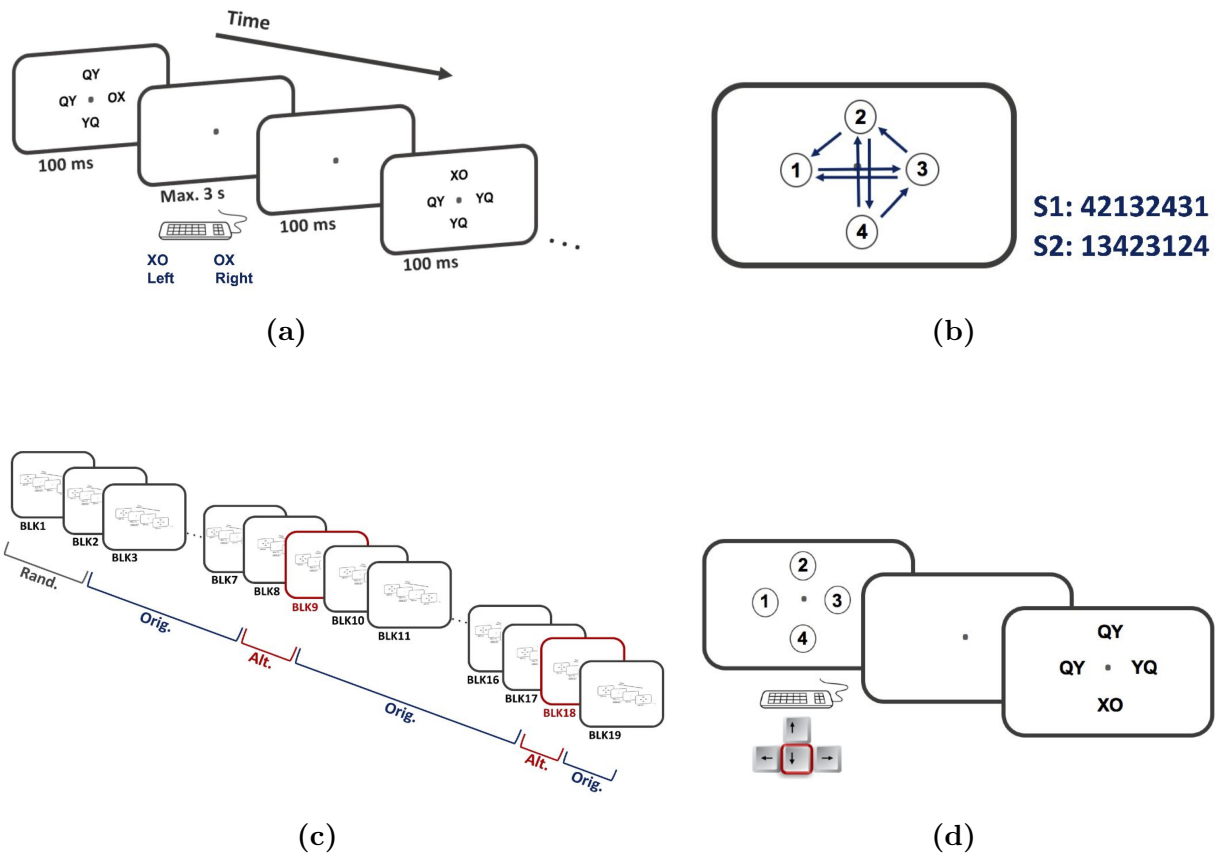


Figure 2.1: Experimental paradigm of the first experiment. (a) Each trial contains four pairs of letters: one target ("XO"/"OX") and three distractors ("YQ"/"QY"). Participants have to respond to the target's identity: "XO" requires a left response, and "OX" requires a right response. Each trial displays for 100 ms, after which only the fixation point remains on the screen until a response is given or the maximum time of 3000 ms elapses. The next trial starts 100 ms after a response is given. Therefore, the response stimulus interval (RSI) is 100 ms. (b) There are four possible locations for the target, with 1 referring to the left of the fixation point (9 o'clock) and the remaining numbers to the locations in a clockwise manner (12, 3, and 6 o'clock). We use two second-order sequences: S1 and S2. Arrows in this figure are only demonstrating the trajectory of the sequence S1 as an example and are not used in the experiment. (c) The task contains a total of nineteenth blocks. The first two practice blocks use random patterns to present the targets, and the remaining blocks, except for the transition blocks (blocks 9 and 18), are organized according to the primary pattern (i.e., the more frequent pattern). In the transition blocks, the primary sequence (S1 or S2) is replaced by the alternative, which is the less frequent sequence (S2 or S1, respectively). (d) During the generation tasks, participants have to choose between four possible locations on the screen with the corresponding four arrow keys on the keyboard. After 400 ms blank, the trial is presented for 600 ms with a target showing in the chosen location.

2.3.0.3 Behavioral paradigm

Experiment 1 The experimental design was developed using the pure perceptual learning (Coomans et al., 2012) and statistical-sequential learning (Summerfield & de Lange, 2014) paradigms. Each trial started with a 1000 ms presentation of a black fixation dot followed by a 100 ms presentation of a stimulus that was centered on the fixation dot. Participants were instructed to respond to the identity of the target ("XO" or "OX") as quickly and as accurately as possible over a delay period with a maximum response time of 3 seconds. During this delay period, only the fixation dot was visible. The target "XO" (where the letter "X" appears on the left of the letter "O") required pressing the button "C" on the keyboard with the left index finger, and the target "OX" (where the letter "X" appears on the right of the letter "O") required pressing the button "N" with the right index finger. Once participants responded, the next trial was started after a 100 ms response-stimulus interval (RSI; see Figure 2.1a). To hinder the oculomotor movements, first, we urged participants to focus on the fixation dot without making eye movements, and second, we used a small visual angle to display the letter pairs ($< 1°$). Each trial was composed of one target and three distractors. The identities of targets and distractors always varied randomly with an equal probability.

The task started with two practice blocks, each consisting of 54 trials, aimed at familiarizing the participants with the experimental procedure. During these practice blocks, the target's locations underwent pseudorandom changes, with the consideration that the same location never appeared consecutively or more than once in succession. Following the first two practice blocks, participants received feedback regarding their error rates. Subsequently, a self-paced break lasting between 0.5 and 5 minutes was provided.

After this initial phase, participants proceeded to complete seventeen experimental blocks, each comprising 104 trials. Following the completion of each block, participants were granted a self-paced break lasting at least 0.5 minutes, with a maximum duration of 5 minutes.

Unbeknownst to the participants, the targets' locations within these blocks were orga-

nized according to a second-order 8-element sequence. Specifically, two sequences (S1 and S2) adopted from Coomans et al. (2012) were utilized, each consisting of the following respective order of the target locations: 42132431 and 13423124 (refer to Figure 2.1b). Each sequence was repeated 13 times in each block. For each participant, one of the sequences was selected to organize the majority of the paired letters' locations, which we refer to as the *primary* sequence. To evaluate the extent to which learning occurred from acquiring knowledge of the underlying sequence versus familiarity with the task, we introduced blocks in which the paired letters were presented in accordance with the *alternative* sequence. These blocks, referred to as the alternative blocks, were presented in block 9 and 18. Having two transition blocks in the first experiment helped us compare the learning trend between the first and second halves of the blocks. This comparison aided us in determining the optimal placement of the transition block for our subsequent experiment.

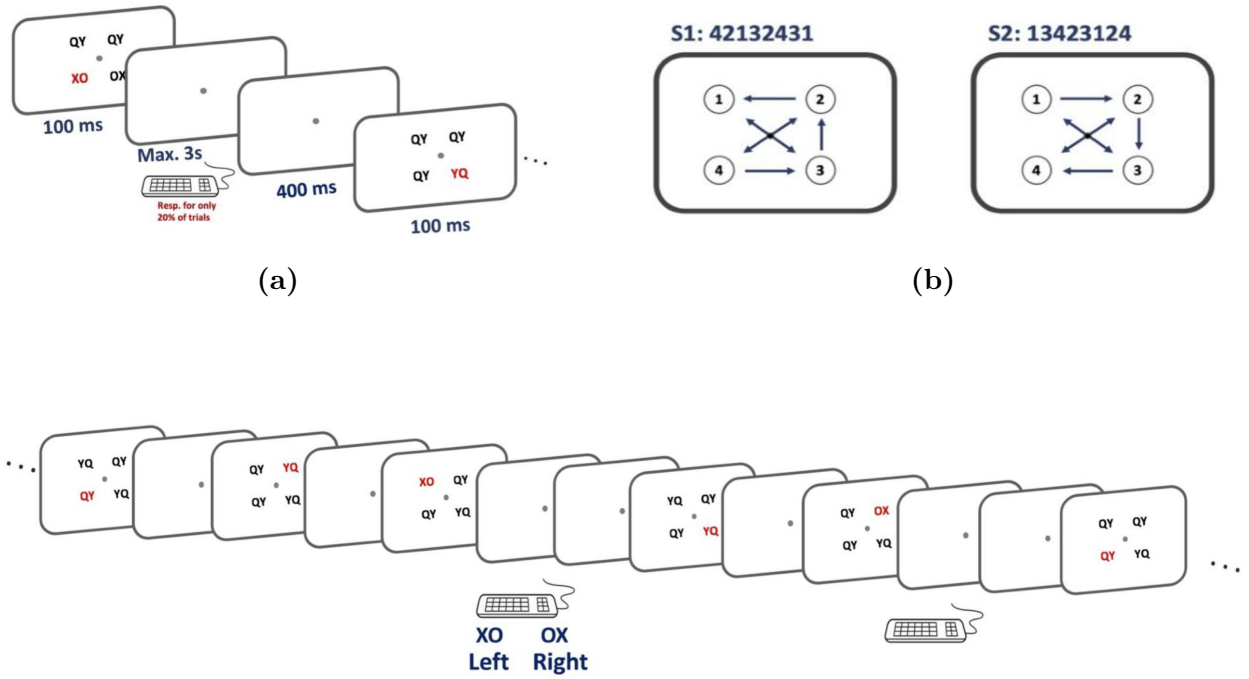


Figure 2.2: Experimental paradigm of the second experiment. (a) Illustrating the sequence of events in each trial, including the presentation of a stimulus for 100 ms, a response time window of up to 3 sec for trials requiring responses, and a subsequent response-stimulus interval (RSI) of 400 ms. (b) The task utilized two second-order featured sequences (S1 and S2). The arrangement of paired letters is based on a square shape. The numbers represent the locations of the paired letters within the square configuration. with location 1 corresponding to the top-left position, and the remaining locations progressing clockwise from that point. The arrows in the figure are only for illustration purposes to show the transition between locations and were not used in the actual experiment. (c) A sequence of trials is provided as an example. In each trial, one of the paired letters is highlighted in red. The locations of these red paired letters follow the arrangement of S1 in this example. Participants are prompted to respond to a trial only when the target ("OX" or "XO") is presented in red. In the example provided, a response is required for the third and fifth stimuli. Specifically, a right response is required for "OX", and a left response is needed for "XO". Each stimulus includes three distractors (YQ" or QY"), with no response necessary when a distractor is presented in red.

Experiment 2 The task in this experiment was designed to address some of the issues we had in the first experiment. In the second experiment, participants performed the task with

similar main logistics used in the first experiment. Similarly, in this experiment, participants were required to do a binary classification task to distinguish between "XO" (left response) and "OX" (right response) by pressing two different buttons on the keyboard. In this iteration, we recorded EEG signals throughout the duration of the experiment, as our focus was to investigate the neural signatures associated with sequential predictive learning.

An initial concern we aimed to address from the first experiment was the arrangement of the paired letters. Initially, we presented our stimuli in a diamond shape in each trial. However, due to the difference between the left and right brain hemispheres, we presented the stimuli in a square shape in the current experiment to reduce any potential confound of the targets' spatial locations. By organizing paired letters in a square shape instead of a diamond shape, half of the letters were on the left and the other half on the right side of the screen. Therefore, we arranged paired letters based on a square shape (Figure 2.2b), where location 1 referred to the top-left corner of the square and the remaining numbers to the locations in a clockwise manner. We used the same two second-order sequences as in the first experiment (Figure 2.2b). Each participant completed 19 blocks after the four practice blocks (23 blocks in total), with blocks 20 and 21 as the transition blocks. According to our findings from the first experiment, we decided to include the alternative blocks in the second half of the blocks. This allowed sufficient learning of the primary sequence before transitioning to the alternative sequence. Additionally, the use of two alternative blocks, rather than one, was implemented to increase the number of trials utilizing the alternative sequence.

Another concern that arose was the potential interruptions due to the need to make many motor responses. Given our focus on achieving pure perceptual learning, the necessity of responding to each trial introduced an extra layer of interruption. Therefore, to reduce these interruptions in experiment 2, the number of trials in which participants were required to respond was reduced to 30%, with the target presented in the color red as the response cue (Figure 2.2c). Participants had to engage sustained attention to all trials in order to detect

the target trials, thus enabling learning on all trials even while only responding on 30% of them.

Furthermore, the response stimulus interval (RSI) was increased from 100 ms to 400 ms (Figure 2.2a) to provide a longer pre-stimulus time window for our EEG processing. Given the increased task complexity, the experimental procedure began with four practice blocks, each consisting of 54 trials to familiarize participants with the task. During the initial two practice blocks, the target's location with the color red changed pseudorandomly, with the condition that the same location never appeared consecutively or twice in succession. All trials during these blocks required a response from the participants to facilitate their familiarity with the task. In contrast, the last two practice blocks were designed to mimic the main experiment, wherein participants were required to respond only 30% of the total trials where the target appeared in red. Following the completion of the four practice blocks, the participants received feedback on their error rates. Only participants who achieved performance above 70% during the four practice blocks were selected to participate in the main experiment.

The main experiment comprised 19 blocks, in addition to the four practice blocks, totaling 23 blocks. Each participant completed a total of 2192 trials throughout the experiment, consisting of 54 trials for each of the four practice blocks and 104 trials for each of the 19 blocks in the main experiment. A self-paced break of at least 0.5 minutes, up to a maximum of 5 minutes, was provided to participants following the completion of each block.

2.3.0.4 Implicit learning

After the completion of the behavioral experiment, a series of post-experimental tasks were conducted to evaluate the extent of participants' implicit learning of the underlying sequence. First, participants were asked to write down their understanding of the task's goal on a designated sheet of paper. Following that, on the same sheet, participants were asked a set of 15 multiple-choice questions specifically designed to assess their awareness of the underlying

sequence. Once the questionnaire was completed, participants were informed that the target locations followed a regular pattern that was repeated throughout the task.

Subsequently, a generation task consisting of inclusion and exclusion versions, as well as a recognition task, were conducted (Destrebecqz & Cleeremans, 2001). In the inclusion generation task, participants were asked to generate 54 series of the target's locations that resembled the primary sequence in the main experiment as much as possible, while ignoring the distinction between "XO" and "OX". Participants were instructed not to generate the same location twice or more in succession and to rely on their intuition when they could not recollect the target's location. In each trial, four possible locations surrounded a fixation point with Arial font size 10, and participants had to choose the target's location by pressing one of the four spatial compatible keys on the keyboard: four arrow keys for the first experiment (Figure 2.1d), and "G", "H", "V", and "N" keys for the second experiment. After participants selected the location, a trial akin to the main experiment was presented on the screen. In this trial, a target (highlighted in red in the second experiment) was positioned at the chosen location, while three distractors were placed in the remaining locations. This generated stimulus was presented for 600 ms, followed by a 400 ms blank period. In the exclusion generation task, participants were instructed to avoid reproducing the sequential regularities of the target's locations with the constraint of not repeating immediate locations. Participants had to generate 54 trials using the same keys as the inclusion task.

Lastly, participants performed a recognition task where a series of 3-length sequences were presented on the screen. Each trial in these sequences was presented for 800 ms, and participants had to determine whether the underlying pattern of the target's locations in the presented 3-length sequence was "familiar" or "unfamiliar" to them. If they believed that the sequence of the target's locations was part of the primary sequence, they had to choose "familiar" by pressing "C" on the keyboard using their left index finger. Otherwise, they had to select "unfamiliar" by pressing "N" on the keyboard using their right index finger. In total, participants responded to 66 questions.

Indeed, all the post-experimental complementary tasks were designed to assess participants' implicit learning of the structural sequence of the target's locations. Specifically, the generation tasks evaluated the level of control participants had over their acquired knowledge, enabling them to generate sequences that either resembled or did not resemble the primary sequence. On the other hand, the recognition task assessed their capacity to recognize any familiar patterns, even if they might not have been able to regenerate the sequence themselves.

	Sequence 1 : 42132431			**Sequence 2 :** 13423124		
	Previous	**Current**	Next	Previous	**Current**	Next
Group 1	2	**1**	3	3	**1**	2
	3	**1**	4	4	**1**	3
Group 2	4	**2**	1	4	**2**	3
	3	**2**	4	1	**2**	4
Group 3	4	**3**	1	2	**3**	1
	1	**3**	2	1	**3**	4
Group 4	1	**4**	2	2	**4**	1
	2	**4**	3	3	**4**	2

Figure 2.3: Illustration of how the trials were organized into groups based on the current target's location within the sequence (S1 or S2). Group 1 denotes the current stimulus occupying position 1, and the same logic is applicable to the remaining three groups. The experimental sequence exhibited a second-order property, signifying that each current location had two potential subsequent locations. For instance, in Group 1, the current target placed at location 1, and the prospective upcoming locations in S1 were 3 or 4, while in S2, they were 2 or 3. Consequently, we established four distinct groups for each sequence, encompassing the combination of current and next locations. This grouping methodology enabled us to explore the influence of predictability regarding the upcoming location with an anticipated chance level of 50%.

2.3.0.5 EEG recording and preprocessing

The Brain Products actiCHamp recording system (Brain Products GmbH) was used to continuously record EEG signals. The recordings were obtained from a wide range of scalp electrodes: FP1, FP2, F3, F4, F7, F8, C3, C4, P3, P4, P5, P6, P7, P8, P9, P10, PO3, PO4, PO7, PO8, O1, O2, Fz, Cz, Pz, POz, and Oz. The electrodes placed on the left and

right mastoids were utilized as reference sites. Horizontal eye movements were detected by placing electrodes laterally to the external canthi, while the vertical EOG was recorded from an electrode below the right eye to detect eye blinks and vertical eye movements. Electrodes' impedance was kept below 50 $K\Omega$. All signals were recorded single-ended and then referenced offline. The EEG signals were filtered online using a cascaded integrator-comb antialiasing filter with a half-power cutoff at 130 Hz, and the signals were digitized at 500 Hz.

The EEG signals were processed and analyzed in MATLAB 2021a using the EEGLAB 2021.0 (Delorme & Makeig, 2004) and ERPLAB 8.30 toolbox (Lopez-Calderon & Luck, 2014). Initially, the scalp EEG was referenced offline to P9. Next, all signals were bandpass filtered using a noncausal Butterworth impulse response function, with half-amplitude cutoffs at 0.1 and 80 Hz and 12 dB/oct roll-off, and were resampled at 250 Hz. We also shifted the data 27 ms to account for the monitor delay. To remove the power line noise, a notch filter at 60 Hz was also applied.

A bipolar horizontal EOG derivation was calculated as the difference between the two horizontal EOG electrodes (left HEOG − right HEOG), and a vertical EOG derivation was computed as the difference between Fp2 and the electrode below the right eye (FP2 − VEOG). The data was then visually inspected to identify and remove the components of the common ocular, cardiac, muscular artifacts, or extreme voltage offsets. We then used independent component analysis (ICA) on the scalp EEG for each participant to identify and remove the components that were associated with blinks and eye movements (Jung et al., 2000; Drisdelle et al., 2017).

After the ICA correction, channel P9 was added, the EEG data was referenced off-line to the average of all electrodes and then re-referenced to the left and right mastoids, and segmented into epochs of -400 ms to 1000 ms relative to the stimulus onset of the current location. Additionally, to remove the epochs and channels containing artifacts, we used the Commonly Recorded Artifactual Potentials (C.R.A.P.) algorithm (Luck, 2022). We then excluded trials from analyses using the simple voltage threshold, with the voltage threshold

equal to 150 μV over the whole epoch for all the channels. We then performed step-like artifact detection using the step function on the vertical EOG derivation (FP2 − VEOG) with the blink threshold of 70 μV over the whole epoch. Next, we applied step-like artifact detection on the horizontal EOG derivation (left HEOG − right HEOG) with the blink threshold equals 32 μV over the whole epoch.

2.3.0.6 Grouping and data preparation

To prepare the data for our analysis, a unique grouping method was employed (Figure 2.3), leveraging the second-order feature of our sequences.

To make the terminology clear in this paper, we used the words 'previous', 'current,' and 'next' to describe the order of the stimuli within the epochs (see Figure 2.3). It can refer to both the trials and the locations of the targets within a trial. For example, 'current trial' refers to the trial presented at the current time (i.e., onset time of 0 ms), and the 'current location' refers to the location that was used to present the target in that specific trial. Similarly, previous trials signify the trials preceding the current trial, and 'next trials' are those following it (columns in Figure 2.3).

As an example, consider sequence 1: if location 1 was chosen to display the target, owing to the second-order nature, two potential locations existed for both the preceding and succeeding targets. Consequently, locations 2 and 3 are considered 'previous locations', and locations 3 and 4 are considered 'next locations'.

In addition to the usage of 'previous', 'current', and 'next,' we further categorized the trials into four groups based on the location of the 'current' trial on the screen (rows in Figure 2.3). Specifically, when the current location was location 1, the corresponding group of trials was labeled as 'group 1'. In this group, we observed two distinct bins: bin 1 represented the chunk `'2 -> 1 -> 3'`, while bin 2 represented the chunk `'3 -> 1 -> 4'` of the sequence. Likewise, when the current location was location 2, the group was identified as 'group 2', and so forth, resulting in a total of eight groups across the two sequences and therefore eight

unique trial types for each sequence (refer to Figure 2.3).

2.3.0.7 Decoding

To measure the extent of decodable information associated with the generation of predictions about future events, we employed a direct and systematic EEG decoding approach (Bae & Luck, 2018; Lopez-Calderon & Luck, 2014; Luck, 2022). This technique aimed to extract and analyze any decodable patterns in the neural EEG signals that could be linked to predictions of upcoming locations within our experimental paradigm. This decoding approach contrasts with the widely used forward encoding methods (Serences et al., 2009; Brouwer & Heeger, 2011; Fahrenfort et al., 2017), which can introduce assumptions about the nature of underlying representation that might not align with specific feature dimensions and are not directly pertinent to the research questions posed in our current study.

Moreover, as part of our exploratory analysis, we conducted a time-frequency analysis to investigate whether different frequency bands carried decodable information pertaining to upcoming stimulus locations. To capture a comprehensive range of neural activity, we considered various frequency bands including alpha, theta, beta, and gamma, covering frequencies spanning from 4 to 48 Hz. We noted comparatively lower decoding accuracy (for additional details, see supplementary figures in Section 2.7).

However, it is important to note that time-frequency analysis has the potential to introduce temporal smearing due to the application of filters to extract distinct frequency components (de Cheveigné & Nelken, 2019). This smearing effect can complicate the preservation of finer temporal details within the analysis. To ensure that we maintained optimal temporal resolution for our investigation, the present study will primarily focus on the decoding results obtained from the ERP analysis.

To focus our analysis on the visual areas that are hypothesized to be associated with predictive learning, we chose 17 channels that were relevant to our study. These channels were P3, P4, P5, P6, P7, P8, P9, P10, PO3, PO4, PO7, PO8, O1, O2, Pz, POz, and Oz. By

analyzing these channels, we were able to gain a more focused and targeted understanding of the neural activity underlying predictive learning in the visual system.

As detailed in section 2.3.0.6, we implemented a grouping strategy by organizing trials according to the current location on the screen. This arrangement led to the creation of two distinct bins, each containing two possible locations for the previous and subsequent trials. This approach enabled us to explore the brain's capacity to predict the upcoming locations throughout the entire epoch, with a chance level of 50% – that is, we used a 2-class classification task using a decoder to extract information pertaining to the prediction of the next locations.

The decoding procedure was performed for each time point independently and for each participant ranging from -400 ms to 1000 ms. To classify the upcoming stimulus location, we used a support vector machine (SVM) with the default kernel of the Matlab fitcsvm() function. The decoding procedure comprised two distinct phases: training and testing. During the training phase, a separate SVM was trained to differentiate between the locations of the upcoming stimuli, and in the testing phase, the trained SVM was used to classify new data that was held out for testing.

To allocate trials for the decoding of each time point, we used a 3-fold cross-validation procedure. First, we divided the trials into three folds, with each fold consisting of equal numbers of trials for each of the two upcoming locations. For each fold, we averaged the trials to produce a scalp distribution for the analyzed time point, which resulted in a matrix of 3 folds $\times$ 2 locations (bins) $\times$ 17 visual electrodes. Two of the three folds were used to train the SVM model with known location labels. Next, the trained SVM model was used to predict the location of the upcoming stimuli for each of the unlabeled locations in the held-out fold that was reserved for testing, using the predict() function in MATLAB. The function assigns a location label for each trial in the held-out fold as the test data, and then the decoding accuracy was calculated by comparing the ground-truth location labels with the predicted labels from the model.

To ensure the stability and reliability of the decoding performance, we repeated the 3-fold cross-validation three times for each analyzed time point, and after all three iterations, all the folds were considered as held-out test data. This helped to minimize the risk of overfitting and increase the generalizability of the results. To further reduce inconsistencies associated with the assignment of trials to folds, we repeated the entire cross-validation procedure 10 times with new random assignments of trials to the three folds. The decoding accuracy was then averaged across the two bins, the three cross-validations, and the 10 iterations, resulting in a decoding percentage for each time point that was based on 60 decoding attempts (2 bins $\times$ 3 cross-validations $\times$ 10 iterations). To minimize noise, the averaged decoding accuracy values were smoothed across time points using a five-point moving window, equivalent to a time window of $\pm$40 ms, which produced a more stable and reliable decoding estimate for each time point.

The precision of the temporal resolution obtained from the entire EEG processing and decoding pipeline was estimated to be approximately $\pm$50 ms. This was determined by using a 400 ms boxcar function through the components of the pipeline that generated low-pass filtering. In this analysis, a substantial value was manually added to the EEG signals of one of the two bins between 150 to 550 ms. This deliberate manipulation aimed to simulate a distinctive pattern with the expectation that the decoding process should accurately identify and classify this pattern, leading to a decoding accuracy of 100%. The decoding pipeline was then applied, and the decoding analysis achieved an accuracy rate of 100% within the time period from 200 to 500 ms, confirming an approximate difference of 50 ms.

To correct for multiple comparisons, we employed the false discovery rate (FDR) correction method (Benjamini & Hochberg, 1995). The FDR correction method controls the expected proportion of false positives among all significant results. To perform FDR correction, we used the fdr_bh function in MATLAB, which employs the Benjamini-Hochberg procedure with a dependency assumption ('dep' method) to adjust the p-values. This method is commonly used in neuroimaging studies and has been shown to provide a good balance be-

tween false-positive and false-negative errors while controlling for the overall false discovery rate.

2.3.0.8 Cross temporal decoding

As an extension of the decoding over time, we utilized cross temporal decoding (King & Dehaene, 2014; King et al., 2014) by assessing the model's ability to accurately predict brain activity patterns at different time points. The primary goal here is to see whether training our model mainly within the pre-stimulus time window achieves substantial decoding accuracies when evaluated on future time points. Hence, our underlying assumption is that if the brain leverages information about pre-stimulus activity patterns and predictive learning, we could potentially observe a substantial decoding accuracy when training on the pre-stimulus time window and subsequently testing on future time points.

To implement cross temporal decoding, we divided our data into distinct time bins (every 20 time samples). We then trained a linear classifier (i.e., SVM) on data from a particular time point to predict the location of the upcoming stimuli. Following training, we tested the model's predictive performance by evaluating its ability to generalize to different time points. Similar to our decoding procedure, we used 3-fold cross-validation. We divided the data at each time point into three folds and used two folds for training at one time point and one fold for test at a different time point. We then repeated that 3 times so all the folds at the other time point would be chosen as a test fold. We performed the whole process for 10 iterations.

By repeating this process across various combinations of training and testing time points, we constructed a cross temporal matrix of size 70×70. Within this matrix, each cell contains the decoding accuracy associated with the specific pairing of training and testing time instants. The visualization of this matrix enabled us to discern the temporal progression of decodable information within our data. To assess the statistical significance of our findings, we performed right-tailed t-tests using MATLAB 'ttest' function. Subsequently, we applied

FDR correction to account for multiple comparisons, identifying data points that showed statistically significant accuracy above the chance level.

Cross-temporal analysis provides a straightforward method to analyze the signatures and patterns linked to predictive learning. Additionally, it enables the investigation of information transferability across diverse temporal contexts.

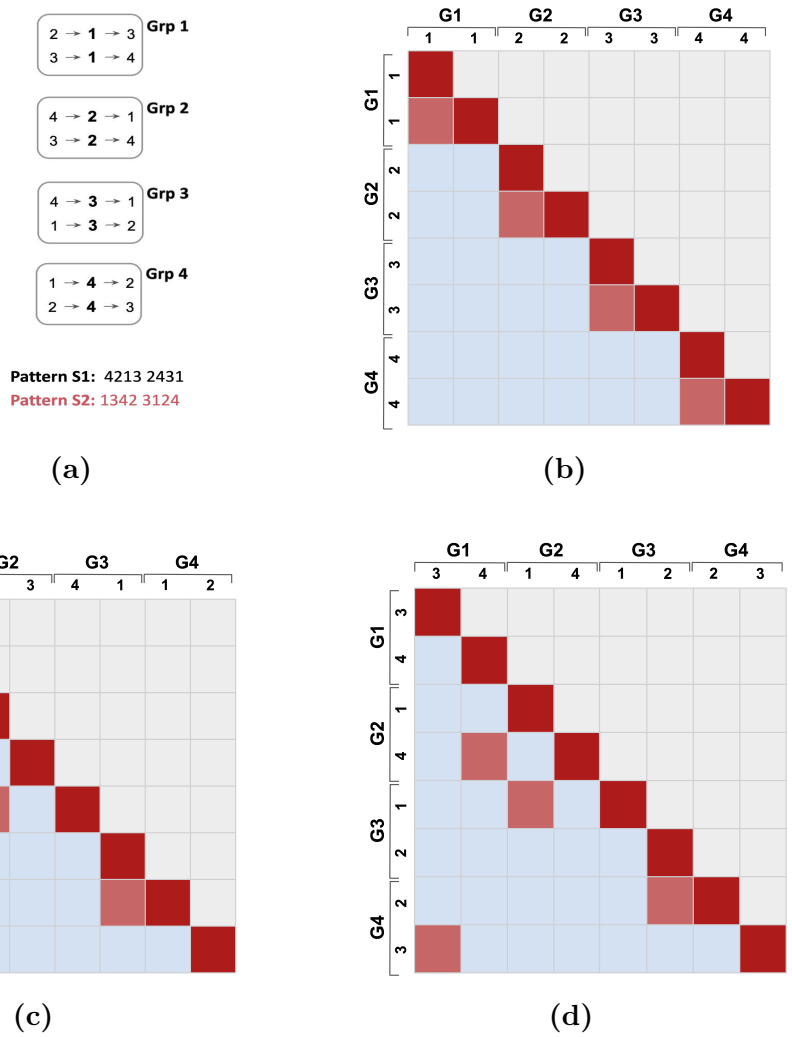


Figure 2.4: Hypothetical RSA matrices for sequence 1. (a) Depicts spatial grouping based on the current location. (b) Illustrates hypothetical RSA for trials at time $t = 0$ ms. The ordering is based on each group's current location. For example, G1 corresponds to group 1, where both bins within that group have location 1 as the current location. (c) Presents hypothetical RSA for preceding trials. The ordering on the top and left side of the matrix is based on the previous locations of the two bins for each group. For instance, group 1 denoted by G1 has locations 2 and 3 as its previous locations. (d) Shows hypothetical RSA for succeeding trials. Similar to the previous hypothetical RSA, but here the ordering is based on the next locations within each group for two bins. For example, G1 for group 1 has locations 3 and 4 as the next locations. All RSAs are symmetric; therefore, we only included the lower triangle for our analysis. The diagonal, denoted by the color red, represents the highest similarity, and it was also excluded from our analysis. Pink-shaded cells within each hypothetical RSA indicate cells expected to have higher similarity.

2.3.0.9 Representational similarity analysis

Decoding techniques have a limitation in that they cannot provide specific information about the decoded data. Therefore, we examined the representation of the decoded information to determine the type of information being processed. To accomplish this, we utilized representational similarity analysis (RSA) (Nili et al., 2014; He et al., 2022), a method that compares the structure of brain activity elicited by different types of trials. In our study, we were interested to know whether the representation of a trial is more similar to the representation of the current trial on the screen or is more similar to the representation of the previous or next locations.

We used our grouping strategy, which organized our trials into four groups, with each group containing two bins (Figure 2.3), with each bin representing the same current location, while differing in their previous or next locations (Figure 2.4a). If the neural response is more similar to the current trial, it suggests that the two bins 'within' each group should exhibit more similar neural responses. However, if there is a higher similarity to the previous or next trial, then the similarity within the group decreases, and there is greater similarity 'across' groups for bins with the same previous or next location.

We utilized RSA to investigate the neural responses induced by the different bins, both within and across groups. Each comparison of these responses is represented by a single cell in our RSA matrix. We then calculated the similarity of the brain activity for each possible pair of cells in this matrix. The term neural response refers to the distribution of voltage values across scalp electrodes. We used the standard Pearson's correlation coefficient to compute the similarity between the voltage distributions on the scalp elicited by different bins. More specifically, we collected the ERP scalp distribution for each bin, whether within or across groups, at each time point. Then, we calculated the correlation between the voltage distributions triggered by all bin pairs.

To generate the representational matrix for each sequence, we used the recorded brain activity over all vision channels for each participant. Each sequence consisted of four groups

with two bins per group (as shown in Figure 2.4a for sequence 1). This gave us a total of eight different cell types, resulting in a final matrix size of 8×8 (see Figure 2.4b) for each sequence and for each subject at each time point. Each cell of the matrix represents the similarity between the neural responses elicited by a given pair. For example, cell (1,3) compares the pattern of brain activity elicited by 'group 1 - bin 1' and 'group 2 - bin 1'. To explore the similarities between different trial types, we compared all pairs of cells. Given that the upper and lower triangles of these matrices mirror each other and that the diagonal values are always 1, we excluded the diagonal and the upper triangle when comparing matrices. To examine whether the information being decoded was more strongly related to the current trial, or the previous/next trial, we built hypothetical RSA matrices for each.

The difference between the hypothetical RSAs lies in the ordering we used, which is indicated on the top and left side of the matrix. In the current hypothetical RSA (Figure 2.4b), we used the current location within each group (e.g., location 1 for both bins in group 1 denoted as G1). While, in the previous hypothetical RSA (Figure 2.4c), the ordering shifted to include the previous locations corresponding to each bin within each group (e.g., locations 2 and 3 in group 1). Similarly, for the next hypothetical RSA (Figure 2.4d), we incorporated the next locations (e.g., locations 3 and 4 in group 1). The anticipated outcome is the highest similarity across groups with the same previous or next location, respectively, for the previous and next hypothetical RSAs. Additionally, the higher similarity is expected within each group that shares the same location for the current hypothetical RSA.

Note that the hypothetical RSA for previous and next trials were the same (i.e., Figure 2.4c and 2.4d have the same cross-group similarity structure), which is a consequence of the second-order nature of the sequences. Thus, we refer to both matrices as the previous/next hypothetical RSA, which can be distinguished from the current hypothetical RSA, but not from each other.

Once we had established the RSA for each participant at every time point and for each sequence, we compared the resulting RSAs with our hypothetical RSAs, using Kendall's τ_A

rank correlation coefficient. This choice was made based on the nature of our hypothetical RSAs, where the models we considered are categorical one-hot vectors that could result in tied values. Kendall's τ_A is a classical rank correlation measure that accounts for ties (Nili et al., 2014).

Additionally, for our statistical test, we utilized the nonparametric Wilcoxon signed-rank to evaluate whether the median correlation coefficient significantly differed from zero. This test is particularly suitable for non-normally distributed data, such as correlation coefficients. We set our alpha level, which determines the threshold for statistical significance, at .05, ensuring a stringent criterion for our statistical analyses. We performed FDR correction to account for multiple comparisons.

Furthermore, to estimate the upper bound of the noise ceiling, we calculated the correlations between individual correlation matrices and the grand average correlation matrix. By taking into account the overall trends observed across participants, this approach allowed us to establish an upper limit for the correlation values within our data. Additionally, to obtain the lower bound of the noise ceiling, we generated separate grand average correlation matrices for each participant, with each matrix excluding the data of the respective participant. This process enabled us to capture the unique contributions of each individual while assessing the lower boundary of the noise ceiling.

Using the RSA approach enables us to compare the similarities between different trial types, allowing us to identify the specific type of information being decoded. This provides valuable insights into the underlying mechanisms of neural processing related to predictive learning and our research questions.

2.4 Results

2.4.1 Behavioral results in experiment 1

To analyze our behavioral data, we first labeled the trials based on the target's locations: "left" (location 1), "top" (location 2), "right" (location 3), and "bottom" (location 4). Subsequently, we segmented blocks 3 to 7 and 11 to 16 into two separate groups and computed the mean reaction times for each (Figure 2.5) to increase the statistical power. However, we treated the block immediately preceding and following the transition blocks individually, enabling us to analyze the changes in reaction times as the more frequent sequence (i.e., primary sequence) shifted to the less frequent sequence (i.e., alternative sequence) in these blocks.

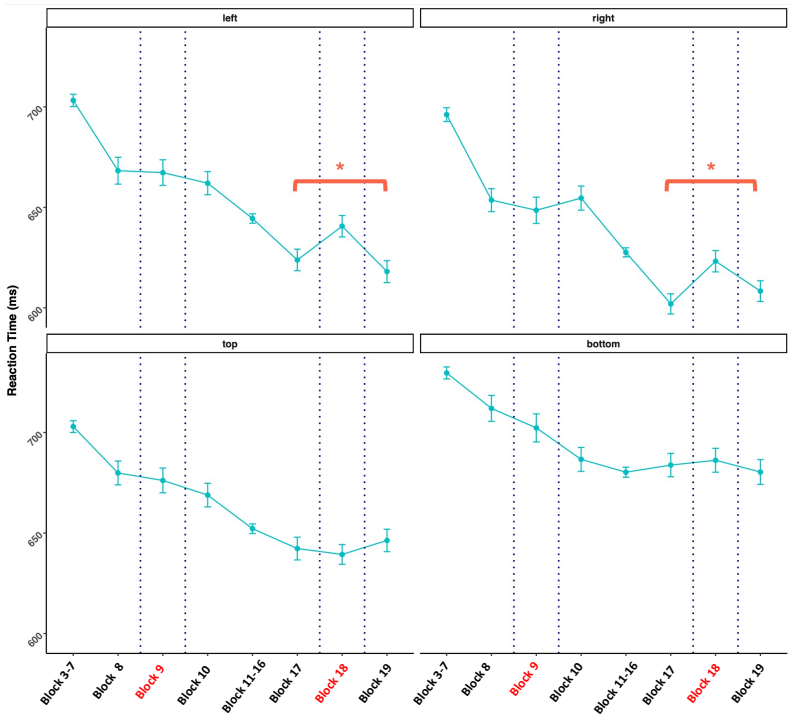


Figure 2.5: Overall performance in terms of speed across blocks in the first experiment. Reaction times are calculated separately for each target's location (left, top, right, and bottom). Reaction times of the training blocks are excluded, so the x-axis starts from the test blocks. Blocks were grouped to enhance the statistical power by increasing the number of trials used for calculating the average reaction times (RTs) within each group. However, blocks before and after the transition blocks, as well as the transition blocks themselves, are considered separately (blocks 8, 9, and 10; and blocks 17, 18, and 19). This approach allows for capturing the change in RTs after transitioning from the consistently repeated primary pattern to the alternative pattern used only in the transition blocks. Stars indicate a significant sequence learning effect, and error bars represent the standard error (SE) of the RTs.

In this experiment, we incorporated two transition blocks with two primary objectives: 1) to compare whether perceptual sequence-based learning occurred early or later in the learning process, and 2) to determine the optimal location for the transition block in the subsequent experiment. Our statistical analyses revealed a consistent decrease in reaction times throughout the learning process for all target locations ($p < 0.05$; refer to Figure 2.5). Interestingly, during the first transition block (i.e., block 9), a continuous decrease in reaction time was observed, despite the shift in the underlying sequence to the alternative one.

Notably, participants responded faster to the trials in comparison to block 8. While a general learning trend was evident from blocks 8 to 10 (characterized by quicker responses over time), this learning might not have solely arisen from sequence-based learning, suggesting that participants might not have fully acquired the underlying sequence at this point.

In contrast, in the second transition block (i.e., block 18), participants' reaction times were influenced by the violation of the underlying statistical pattern, leading to slower reaction times compared to blocks 17 and 19. This observed pattern – an increase in reaction times during the transition from block 17 to 18, followed by a subsequent decrease as the pattern reverted to the primary sequence from block 18 to 19 – provides compelling evidence for sequential learning that occurred in the later blocks. Also, the fact that significant sequence-specific learning was only evident in the left and right locations motivated the use of a square-shaped display in experiment 2 instead of the diamond-shaped one in this experiment.

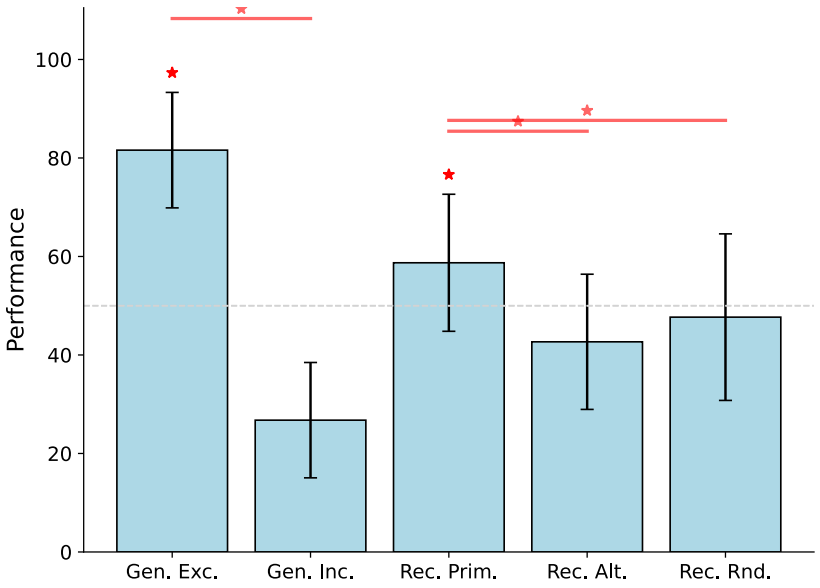


Figure 2.6: Box plot illustrating participants' scores in the implicit learning task of the first experiment. This includes the inclusion and exclusion tasks in the generation phase, along with recognition tasks for the primary, alternative, and random sequences. The outcomes suggest that participants were unable to regenerate the primary sequence in the generation task, yet they exhibited recognition of these sequences, as evidenced by their scores in generation inclusion and recognition of the primary sequence. Additionally, although the recognition of random patterns outperforms the alternative pattern, the difference between these two is not statistically significant.

2.4.1.1 Implicit learning results

Initially, we analyzed the written responses obtained from participants filling out the questionnaires. The results revealed that none of the participants reported being aware of the presence of any underlying sequence within the task.

We then conducted a further analysis of the results from our additional post-experimental tasks. Figure 2.6 displays the outcomes of the two *generation* tasks: inclusion and exclusion, alongside the *recognition* tasks: recognizing the primary, alternative, and random sequences.

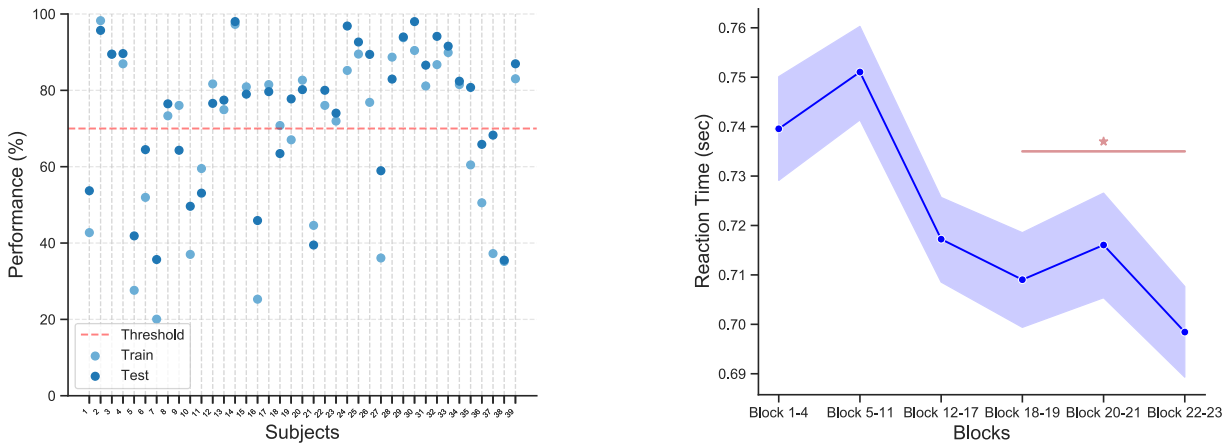
The rationale behind this analysis was based on the idea that if sequential knowledge had been implicitly absorbed during the task, it should naturally impact recognition performance. In other words, The brain's ability to recognize a familiar pattern even without consciously

being able to regenerate it suggests the presence of implicit learning. As shown in Figure 2.6, participants exhibited notably higher accuracy in recognizing the primary sequences in comparison to the alternative and random sequences. Moreover, the recognition performance surpassed the generation performance significantly.

This finding served as an additional confirmation, further reinforcing the notion that participants had indeed acquired the underlying statistical structure of the sequence through implicit learning.

2.4.2 Behavioral results in experiment 2

Our behavioral analysis included a total of 39 participants, of whom 25 participants demonstrated proficient learning in the task (Figure 2.7a). Specifically, these participants achieved an average performance of 70% or higher during the test session. Consequently, the analysis of reaction times (Figure 2.7b) focused exclusively on these 25 subjects.



(a) **(b)**

Figure 2.7: (a) Scatter plot showing the performance of each subject in both the train and test sessions. The x-axis represents the subjects, while the y-axis represents the performance. The red horizontal dashed line indicates the performance threshold (70%) used to select participants for further analysis. Only 25 subjects who achieved a test performance above this threshold were included. (b) Line plot depicting the reaction times (RTs) of the selected subjects across blocks. Participants showed improvement in responding to the target over time, as evidenced by faster RTs and a decrease in overall reaction time across blocks. Notably, during blocks 20 and 21, when the underlying primary sequence transitioned to the alternative sequence, there was a significant increase in RTs. This finding suggests that participants were engaging in sequential learning of the primary sequence. The asterisks indicate the statistical significance of the quadratic pattern from blocks 18 to 23.

In our behavioral analysis, we focused on response trials. Subsequently, we employed mixed linear regression to explore the connection between the reaction times of these trials and the progression of blocks. We incorporated subject variability in all our models. For this analysis, we used "mixedlm" class from the "statsmodels.formula.api" module in the Statsmodels library (Seabold & Perktold, 2010). The outcome of our analysis revealed a significant and negative relationship between RTs and blocks 5-19 ($\beta = -20.351, p < 0.001$), signifying a substantial reduction in RTs as participants advanced through the blocks (depicted in Figure 2.7b). We termed this observed trend as *general* learning, as it seemed to be linked with generalized enhanced task performance rather than explicit acquisition of the underlying sequential pattern.

Furthermore, when encompassing all test blocks (blocks 5 to 23) in our consideration, our analysis yielded results consistent with those observed when including test blocks 5-19. A significant negative correlation between RTs and blocks ($\beta = -9.863, p < 0.001$) emerged, emphasizing a consistent tendency for RTs to decrease as the experiment unfolded. This consistent pattern further supports the idea of progressive performance improvement and learning across the entire experiment.

Next, we quantify the degree to which this learning is attributable to acquiring knowledge of the underlying sequence by comparing the RTs for trials presented according to the primary sequence, which predominated during the majority of the blocks, with those organized according to the alternative sequence introduced in blocks 20 and 21. If the learning trend were solely due to task improvement, we would expect RTs to continue decreasing even when the sequence changed in blocks 20 and 21. However, the results demonstrated that participants exhibited an increase in RTs during these transition blocks, indicating a disruption in performance (Figure 2.7b). Interestingly, when the sequence reverted back to the primary sequence in blocks 22 and 23, RTs decreased again. This fluctuation of increasing and decreasing RTs in blocks 18 to 23 provided compelling evidence for sequential learning (Deroost & Soetens, 2006).

To conduct a statistical analysis on this observation, we focused our analysis on blocks 18 to 23. Within this subset of blocks, we discovered a more intricate relationship between blocks and RTs. In addition to the linear effect of blocks ($\beta = 97.044, p = 0.044$), we incorporated a quadratic term (time squared) into the model. The quadratic term exhibited a significant negative coefficient ($\beta = -10.312, p = 0.032$), indicating a curvilinear relationship. This suggests a significant pattern characterized by an initial increase in RTs followed by a subsequent decrease as the blocks progress.

2.4.2.1 Implicit learning results

As described in the section 2.3.0.4, participants engaged in additional tasks at the end of the experiment to assess the degree to which they acquired the underlying sequences implicitly.

First, we analyzed the written form of the implicit learning test. Our analysis revealed that none of the participants reported the presence of any underlying sequence in the task. This finding confirms that the subjects acquired the statistical structure of the sequence without conscious awareness.

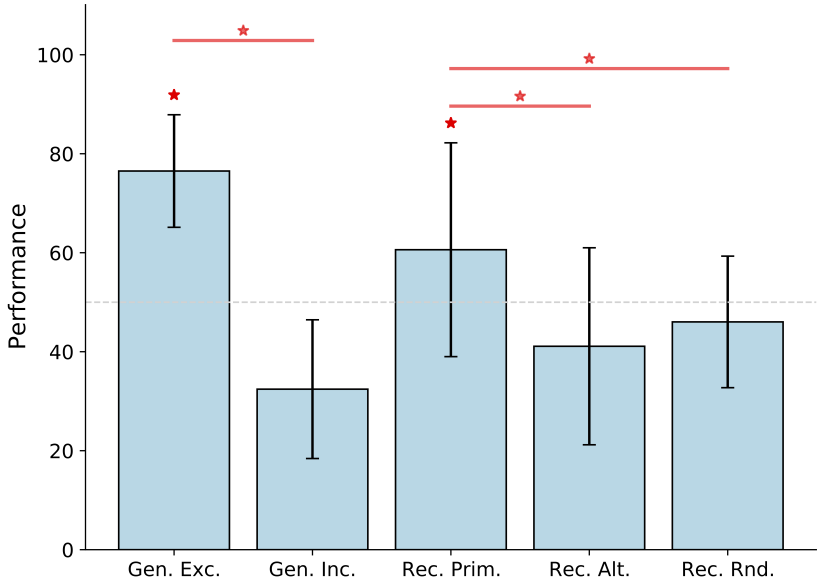


Figure 2.8: Box plots illustrating the implicit learning scores for the two generation tasks conducted in the second experiment—exclusion and inclusion—as well as the recognition of various sequences: primary, alternative, and random sequences. The findings indicate a higher level of proficiency in recognizing the primary sequence compared to the alternative and random sequences. Furthermore, participants displayed a lack of control over their knowledge to regenerate the underlying pattern, as evidenced by their generation inclusion performance. Although the recognition of random patterns outperforms the alternative pattern, the difference between the two is not statistically significant.

Furthermore, we examined the results of the generation and recognition tasks to further explore the participants' implicit learning of the sequential pattern. Figure 2.8 presents the outcomes of the two generation tasks: inclusion (Gen. Inc.) and exclusion (Gen. Exc.), as well as the recognition task for our sequences: primary (Rec. Prim.), alternative (Rec.

Alt.), and random sequence (Rec. Rnd.). The analysis revealed that participants performed significantly better than chance in the exclusion task ($p < 0.001$) and there was a significant decrease in the number of correctly generated chunks from the exclusion to the inclusion instructions ($p < 0.001$). These findings suggest a lack of conscious control over the knowledge of the sequence necessary to regenerate the underlying pattern, which provides support for the presence of implicit learning.

Additionally, we explored whether participants would exhibit recognition of the sequence even if they were unable to explicitly generate it. The aim of this analysis was to determine if the acquired sequential knowledge would automatically influence performance during the recognition task. To investigate recognition performance, we quantified the number of correctly recognized chunks in the primary, alternative, and random sequences. The results, shown in Figure 2.8, indicate that participants significantly recognized the primary sequences above 50% ($p < 0.05$). Furthermore, performance in recognizing the primary sequence was significantly better than recognizing the alternative sequence ($p < 0.05$) and the random sequence ($p < 0.05$). These findings provide additional support for the participants' implicit acquisition of the underlying sequential knowledge. It is also important to note that even though the performance in recognizing the random sequences appears to be greater than that of recognizing the alternative sequences, the difference is not statistically significant ($p > 0.1$).

In conclusion, the participants' lack of awareness regarding the underlying sequence, as evident from their responses to the questionnaire, coupled with their inability to consciously control the sequence knowledge required for regenerating the pattern, and their proficiency in recognizing the familiar sequence pattern, strongly indicates the prevalence of implicit learning among the participants.

In summary, our findings indicate that participants effectively learned the task, and this learning can be attributed to acquiring knowledge of the underlying sequence through perceptual observation. Notably, this learning occurred implicitly, without participants being

consciously aware of the underlying pattern while learning the sequence. Building upon these promising results, our next step is to delve deeper into the implications of these findings by focusing primarily on analyzing the neural EEG signals.

2.4.3 EEG results in experiment 2

2.4.3.1 Decoding results

To investigate the presence of predictive learning used for predicting upcoming locations, we tested whether a decoder trained on ERP signals across a time window centered on the current trial could accurately predict the identity of the next upcoming location. As detailed in the methods, this was done using a cross-validation method with 3 random train / test folds across trials within each subject over 10 iterations, with the False Discovery Rate (FDR) correction method applied to all p values to account for multiple comparisons.

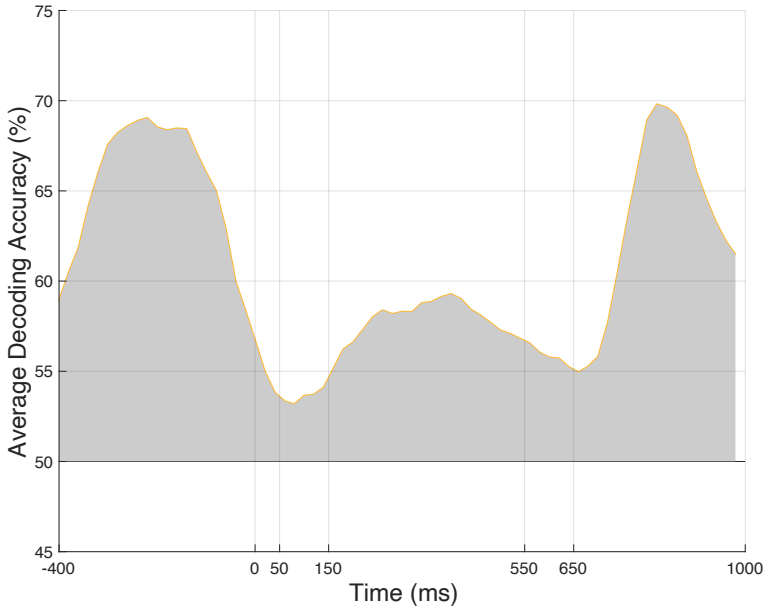


Figure 2.9: The ERP decoding accuracies, with a baseline period of [0:100] ms, consistently remained above chance levels throughout the entire 1400 ms epoch. The shaded region indicates time points that are statistically significant, determined using FDR correction. The vertical dashed lines correspond to key temporal events: the onset time (time = 0 ms), the time interval 50:150 ms during the processing of the current trial, considering a potential 50 ms delay for information propagation to reach the cortex, and the time intervals between 550 ms and 650 ms correspond to the processing of the next locations, accounting for the 50 ms shift.

The results, as illustrated in Figure 2.9, displayed a notable increase starting from the beginning of the epoch (time = -400 ms), surpassing the chance level with a highly significant p-value of $p < 10^{-4}$. Furthermore, these accuracies consistently remained above chance (Figure 2.9, shaded region), as confirmed by the mean of significant t-values and adjusted p-values for our subjects across four groups ($t(95) = 8.9532, p = 0.00001; \text{one} - \text{sample}, \text{t} - \text{test}$).

Considering the systematic grouping of trials during data preparation for the decoder, it's important to note that there exist two potential previous locations (e.g., loc 2 and 3 in group 1) and two potential next locations (e.g., loc 3 and 4 in group 1) presented within the time intervals of [-500:-400] ms and [500:600] ms across the two bins, alongside the same location being shared across both bins (e.g., loc 1 for group 1). In our analysis, as previously mentioned, we consistently account for a 50 ms time delay to accommodate the propagation

of information.

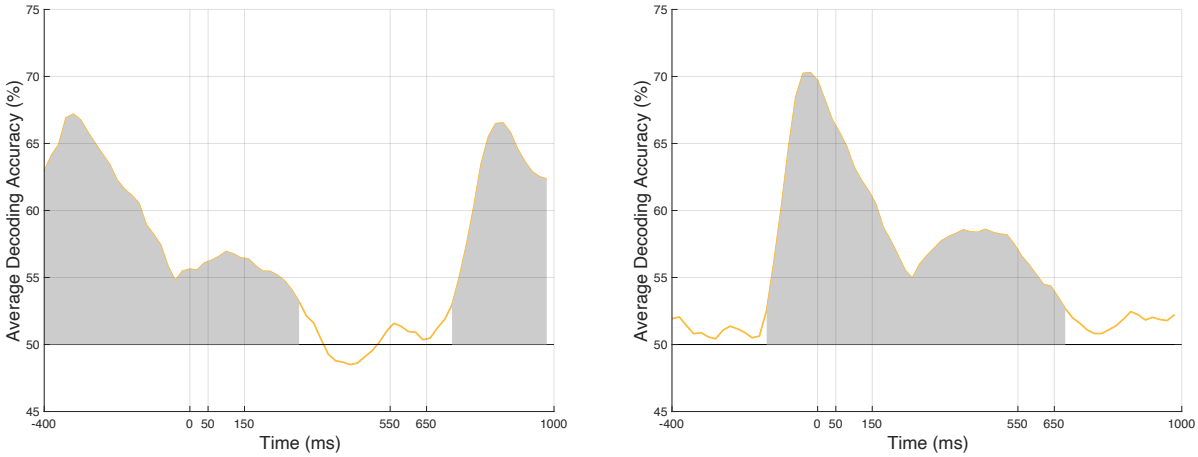
Our analysis revealed significantly high decoding accuracies during the time intervals of [-400:50] ms and [650:1000] ms, along with two prominent accuracy peaks. One possible explanation for the heightened decoding accuracy observed during the [-400:50] ms interval is that the brain retained information from the previous trials prior to processing the current trial. Essentially, the decoding algorithm effectively captured the distinct neural patterns associated with the two potential previous locations, leading to decoding accuracy significantly surpassing chance levels. Notably, a peak accuracy was observed at time = -220 ms, demonstrating an accuracy of 69.04% and a highly significant p-value of $p < 10^{-4}$.

It is important to mention that even though our decoder showed high decoding accuracy in predicting the next locations in our sequences, it is impossible to differentiate this ability from decoding (remembering) the previous locations. The presence of the second-order characteristic in our sequences means that predicting the next location cannot be successfully achieved using solely information from the previous locations; instead, it requires information from two prior locations (second-order information) to successfully predict the next location. This also implies that distinguishing the representation of the previous trial from that of the next trial is impossible, as they are always connected despite being different locations.

As a kind of sanity check on our decoding method, we consistently observed significant decoding accuracy above chance within the [650:1000] ms interval, with a specific peak at time = 820 ms, exhibiting an accuracy of 69.79% and a highly significant p-value of $p < 10^{-4}$. This time window corresponds to when the actual next location is presented on the screen, and thus one would expect to be able to decode these locations with high accuracy. During both of these time intervals, the high decoding accuracy highlights the brain's capacity to distinguish and capture the unique neural patterns linked to the preceding or subsequent locations.

Critically, we observed significant decoding accuracy during the time interval [50:150] ms, when the same current location was present (Figure 2.9). This suggests that the brain has

implicitly learned the underlying sequence and is carrying this predictive knowledge across the ambiguous current stimulus, to be able to effectively predict the next location.



(a) **(b)**

Figure 2.10: Decoding accuracies of ERP signals using different baseline correction approaches. Subfigure (a) illustrates the decoding accuracies without any baseline correction, while subfigure (b) displays the results after applying baseline correction with the [-700:-500] ms interval as our chosen baseline. Both plots demonstrate the robustness of our findings. They clearly depict the presence of decodable information throughout our epoch, suggesting that the choice of baseline correction method does not significantly impact the reliability of our results.

To ensure the reliability of our findings, we did some extra analyses to account for confounding factors that could influence our findings. Specifically, we performed the same decoding analysis using different baseline correction approaches: one without any baseline and another with a baseline set for the time interval [-700:-500] ms. It is important to clarify that the goal of these supplementary analyses was solely to confirm the presence of decodable information and to verify that our choice of baseline was not the primary factor driving the observed results. However, due to the potential for increased noise in the signals, we refrained from conducting an in-depth analysis of these supplementary results.

In the first analysis, we performed the decoding analysis without applying any baseline correction. Surprisingly, even without baseline correction, the decoding accuracy remained significantly above the chance level (refer to Figure 2.10a). When we calculated the mean

of significant t-values and adjusted p-values (after FDR correction) for our subjects across the four groups, the decoding accuracy continued to be statistically significant ($t(95) = 7.5571, p = 0.00132; \text{one} - \text{sample}, \text{t} - \text{test}$). And, it reached its peak at 67.19% with an adjusted p-value of $p < 10^{-4}$. This outcome supports the robustness of our findings and suggests that the presence of decodable information is not solely attributed to our choice of baseline correction.

Additionally, we introduced an alternative baseline timeframe [-700:-500] ms, which corresponds to the blank period before the presentation of the preceding trial. This interval was selected because the locations of the trials presented immediately before the preceding trials (at times -1000 ms to -900 ms) were consistent across the two bins in each group. This provided us with an alternative baseline option during which the trial's location remained the same across the two bins.

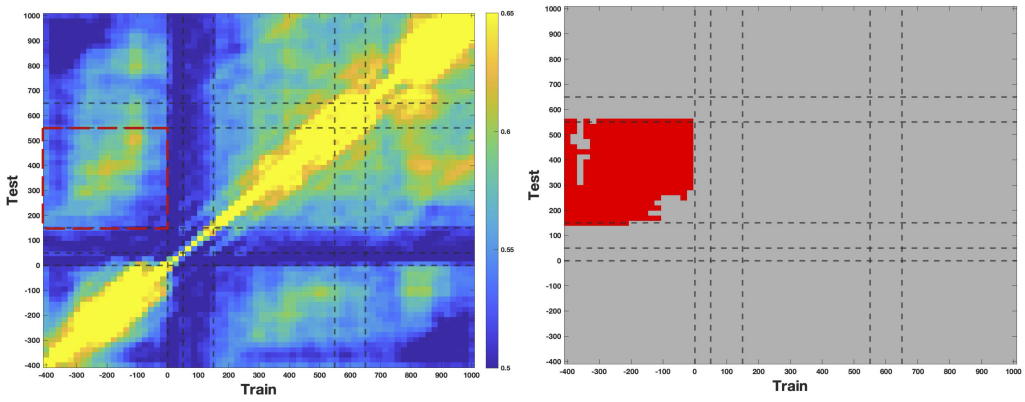
For instance, consider S1 group 1, where at time = 0ms, the trial on the screen was presented at location 1. In this case, the two bins for this group are as follows: bin 1: `'2 -> 1 -> 3'` and bin 2: `'3 -> 1 -> 4'`. Now, if we include one more location, corresponding to the trials immediately before the preceding trials (2 and 3), the sequences for each bin transform as follows: bin 1: `'4 -> 2 -> 1 -> 3'` and bin 2: `'4 -> 3 -> 1 -> 4'`. Notably, location 4 is identical in both bins.

The outcome of adopting this baseline is shown in Figure 2.10b. The results demonstrate that even when selecting a baseline considerably distant from our onset, which introduces more noise compared to our original [0:100] ms baseline, we achieved a notably significant decoding accuracy ($t(95) = 8.4058, p = 0.00234; \text{one} - \text{sample}, \text{t} - \text{test}$). And, it reached its peak at 70.28% with an adjusted p-value of $p < 10^{-4}$.

Both of these outcomes provide strong confirmation of the reliability of our findings, and they collectively emphasize that the choice of baseline did not exert a significant impact on the final results. As a result, we proceeded to perform additional analyses with signals after baseline correction, utilizing the [0:100] ms as the time interval for our baseline.

In summary, our decoding analysis revealed a significant presence of decodable information related to both previous and next locations. Notably, the consistent above-chance decoding accuracy even during the processing of the current trial with the same location across the two bins suggests that the brain has learned the second-order pattern within the sequence to accurately predict subsequent locations, consistent with the behavioral results showing sensitivity to the learned sequences. This demonstrates the brain's capability to capture the predictive relationships embedded within the sequence.

In our subsequent analysis, we delved deeper into the connections between pre- and post-stimulus information, exploring whether past information holds meaningful insights about the information of future time points.



(a) **(b)**

Figure 2.11: Cross temporal results across the entire epoch time. (a) The cross-temporal matrix, sized 70×70, presents the decoding accuracy for various combinations of training and testing time points. Each cell displays the accuracy attained on the held-out test data for the corresponding training and testing time points. Dashed horizontal and vertical lines show event transitions in our epoch. The onset (at time = 0 ms) is when the current location was on the screen. Additionally, we considered a 50 ms time span to account for potential delays in information reaching the cortex. The highlighted red dashed square includes the time window of interest. Within this window, the machine learning model was trained on the pre-stimulus time frame and tested on the post-stimulus time frame. As indicated by the results, many cells showed high decoding accuracy. To establish the statistical significance of these outcomes, a t-test was subsequently conducted, followed by FDR correction for multiple comparisons. (b) Illustrates the outcome of the statistical analysis. Time points other than those included in the selected time frame are shaded in gray, as they were excluded from the statistical analysis. The red data points indicate time points where the decoding accuracy was significantly above the chance level after FDR correction.

2.4.3.2 Cross temporal analysis results

To provide further tests for the existence of predictive learning, an additional analysis was conducted utilizing decoding techniques, where training and testing were performed at different time points. The outcome is a cross-temporal matrix of dimensions 70×70, with each cell representing the resulting decoding accuracy for a given training and test time pairing. The findings for the baseline [0:100] ms are depicted in Figure 2.11a. The x-axis indicates the training time, while the y-axis illustrates the test time points.

Within this matrix, a specific time window of interest, associated with predictive learning, is identified by a red dashed rectangle. In this case, the model was trained on the pre-stimulus time window [-400:0] ms and subsequently tested on the post-stimulus [150:550] ms. The reason for choosing this time window is based on the idea that predictive learning might show up when a model uses information from before the current trial to make predictions about the future, particularly the prediction of the subsequent trial, consistent with the second-order structure.

From Figure 2.11a, it is evident that within this chosen time window, there is considerable decoding accuracy. After performing statistical analysis with adjustments for the false discovery rate (FDR), as shown in Figure 2.11b, we found that the high decoding accuracy during this time window is mostly significant ($p < 10^{-4}$). This finding provides additional support for the presence of a predictive learning pattern.

In summary, the results obtained from our cross-temporal matrix analysis further support the hypothesis that the brain has learned the predictive information present in our sequences, even when not explicitly required for task performance. This suggests that the brain has automatic predictive learning mechanisms in the perceptual system, consistent with the predictive learning theories reviewed earlier. In the subsequent analysis, we focused on the representations underlying the significant decoding found here, to achieve a clearer understanding of the overall mechanisms associated with predictive learning.

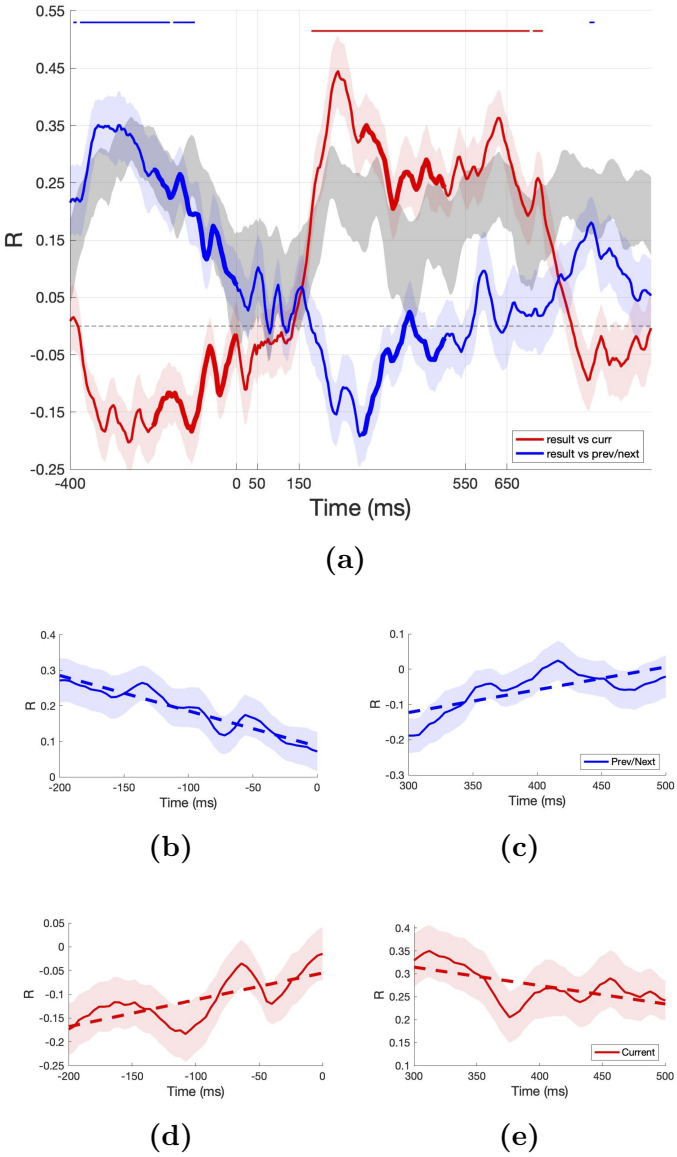


Figure 2.12: Results of the RSA analysis. (a) The comparison between the RSA of each subject and two hypothetical RSAs. Shaded gray areas represent noise ceiling bounds, and horizontal lines on the top indicate significant time points after FDR correction. Subfigures (b) and (d) show comparisons during the pre-stimulus time window [-200:0] ms with previous/next and current hypothetical RSAs, respectively. These zoomed-in versions of the results illustrate an enhancement in similarity to the current trial as we approach the presentation of the current trial. Conversely, there is a decrease in similarity with the representations of the previous/next trials. Subfigures (c) and (e) display comparisons before the presentation of the next trial during the time interval [300:500] ms, with previous/next and current hypothetical RSAs, respectively. This observation shows that as we approach the presentation of the upcoming locations, there is a recovery in the information related to previous/next trials, while the similarity to the current trial decreases. The dashed line indicates the fitted line. Notably, all decreases and increases shown in subfigures (b) to (e) are statistically significant.

2.4.3.3 Representational similarity analysis results

Our decoding results have unveiled a significant amount of decodable information associated with predictive learning of the second-order sequences. To gain a deeper understanding of the nature of this information, we conducted further analyses using the representational similarity analysis (RSA) approach. We constructed representational similarity matrices for each individual subject to compare them against our hypothetical RSAs for current-trial representations (Figure 2.4b) versus previous and next trial representations (Figure 2.4c), 2.4d). This comparative analysis was conducted at each time point and was performed separately for each subject. The aggregated results across all subjects are presented in Figure 2.12a.

Within this plot, each line corresponds to the values of Kendall's τ_A rank correlation coefficient, which indicates the degree of similarity between the subjects' RSA and each of the hypothetical RSAs. More specifically, the blue line shows the comparison with the previous/next RSAs, while the red line represents the comparison with the current hypothetical RSAs. The shaded gray areas highlight the range of the noise ceiling, and statistically significant time points are denoted by horizontal lines positioned above the plot. These significant time points have been identified using FDR correction to account for multiple comparisons.

Furthermore, for a closer examination, we have extracted and presented specific segments within Figure 2.12a. These segments are represented in subfigures 2.12b to 2.12e, and they have been highlighted with bold markings in the Figure 2.12a.

The findings indicate an interesting pattern: initially, before presenting the current trial at time=0ms, we observed a significant similarity with the previous/next RSA. This similarity persisted throughout the blank period until after the current trial's presentation. As we approached the presentation of the current trial, the similarity with the previous/next RSA gradually decreased, while the similarity with the current RSA increased. This trend is particularly evident in the zoomed-in view of the [-200:0] ms time window shown in Figure 2.12b and 2.12d, respectively, where the fitted line demonstrates a significant slope decrease

$(p - value < 10^{-4})$.

Subsequent to the time it takes for the information of the current trial to potentially reach the cortex (typically around 50 ms), we observed a noticeable rise in similarity with the current RSA. As we approached the presentation of the subsequent stimulus at time=500 ms, we noted a decline in similarity with the current RSA, as displayed in Figure 2.12e. Simultaneously, there was a significant increase in similarity with the previous/next RSA, as evidenced in Figure 2.12c, with the fitted line showing a substantial slope increase ($p - value < 10^{-4}$). These changes in representational similarity anticipated the onset of the next stimuli in the sequence, consistent with the hypothesis that the brain has learned the sequences.

As the information from the next trial potentially reached higher-level brain areas, we observed an increase in the similarity with the previous/next RSA. However, it is important to highlight that a high similarity with the current RSA persisted for an extended period of time. This finding aligns with our behavioral results (Figure 2.7), which showed the presence of sequential learning in our data. It is noteworthy that effective utilization of predictive relationships embedded within the sequence to predict the location of the upcoming trial requires the knowledge of the two preceding locations. This observation aligns with the pattern we observe in this result, potentially explaining the prolonged retention of information from the current trial.

Another notable finding is the substantial negative correlation evident between the two types of comparison, depicted by the blue and red lines (pearson correlation coefficient $R = -0.85$, $p < 10^{-4}$). This interesting result may be relevant to an earlier study that emphasized the retention of orthogonal information within distinct subspaces, specific to various cognitive processes like working memory and attention (Panichello & Buschman, 2021). In our context, the brain requires both previous and current location information to successfully predict the subsequent location, but presumably, this information must also be kept distinct. Furthermore, the current stimulus naturally will tend to dominate the

representational similarity structure. Thus, the negative correlation likely arises due to the orthogonal representation of current vs. previous / next locations within different subspaces. While this notion might require further research and experimentation for confirmation, it offers a novel perspective for interpreting these results.

2.5 Discussion

The ability to generalize to novel situations is considered a hallmark of human intelligence. This capacity relies on the emergence of high-level abstract knowledge from lower-level representations, serving as a strong foundation for effectively tackling previously unseen challenges (Behrens et al., 2018; Vikbladh et al., 2019).

In our study, we employed perceptual and statistical learning paradigms across a series of experiments and recorded EEG neural signals to explore the underlying learning mechanisms leading to the formation of such abstract representations, known as predictive learning. To analyze the EEG data, we employed a variety of analytical tools, including advanced machine learning techniques. Our initial analysis focused on decoding the EEG signals (Bae & Luck, 2018; Luck, 2022) to identify any decodable information relevant to our sequences. Subsequently, we expanded this decoding approach to create a cross-temporal matrix (King & Dehaene, 2014; King et al., 2014). Furthermore, we delved into the representations formed using the representational similarity analysis (RSA) technique (Nili et al., 2014).

In summary, our comprehensive analysis of behavioral data and neural EEG signals revealed multiple significant signatures of successful predictive learning, based on the second-order structure of the sequences used. Our behavioral findings demonstrated that participants exhibited both a general speed-up in learning the task and a specific speed-up for the particular sequence they learned. Importantly, this learning occurred implicitly in both of our experiments. The initial decoding findings showed there was significant decodable information about the next, upcoming stimulus location in the sequence. To establish a

stronger connection between past and future information, we then employed a generalized form of the decoding analysis to construct a cross-temporal matrix containing decoding accuracy information for all pairs of training and test time points. The results showed that when we tested a decoder during the post-stimulus time window, which was initially trained on pre-stimulus data, we observed a notably high decoding accuracy. Both decoding-based findings demonstrated that the brain effectively learned to predict the reliable sequences of locations, even though participants were not explicitly instructed to do so. The RSA results further validated that the representational similarity of previous, current, and next information evolved over time with an indication of predictive transitions of representational similarity in anticipation of the upcoming stimulus.

In our decoding analysis, we employed a separate decoder at each time point, consistent with previous research (Bae & Luck, 2018). It focused solely on whether there was decodable information related to upcoming locations, without considering potential connections between past and future information. This analysis primarily served as an exploratory tool rather than a method to precisely determine timing and also the interaction between the previous and future time points. Thus, we used the additional cross-temporal matrix to examine the temporal connection information between the past and future (King & Dehaene, 2014). Interestingly, the cross-temporal results highlighted that within our selected time frame, higher decoding accuracy emerged around 200 ms (i.e., two alpha cycles) before the onset presentation.

In predictive learning, two distinct states are involved: predictions followed by outcomes. It has been suggested that these two states unfold across approximately 100 ms—a timespan often associated with the alpha frequency of 10 Hz (O'Reilly et al., 2021b, 2014). This alpha cycle comprises a 75 ms prediction phase and a subsequent 25 ms for processing real-world input and reducing prediction errors. While previous research has explored different frequency bands, including alpha and theta, in tasks related to attention and predictive processing (VanRullen & Koch, 2003; Mathewson et al., 2009; Min & Park, 2010; Samaha

et al., 2015; Toosi et al., 2017; Fiebelkorn & Kastner, 2019). However, many of these studies include both attention and prediction factors, posing a challenge in isolating the precise contribution of the alpha frequency to predictive processing. Additionally, investigating whether predictions are generated within the initial 75 ms following the 25 ms period in the predictive learning cycle is a complex task. This short and precise temporal window poses a challenge for precise determination that remains unanswered in these studies, as well as in our own, leaving it as an interesting question for future research.

An additional limitation when employing time-frequency analysis and applying filters to extract frequency bands pertains to the potential risk of information leakage over time (de Cheveigné & Nelken, 2019). Such leakage could result in the presence of information from the future in the past time frame that could erroneously be associated with predictive learning. To reduce potential risks involved in EEG processing, having long time windows is recommended, as seen in tasks involving working memory and decoding (Bae & Luck, 2018; Luck, 2022). However, our experimental design presented challenges in this regard. Our hypothetical assumption that learning occurs rapidly through a chain of predictions followed by outcomes every 100 ms influenced our choice of a 100 ms interstimulus interval (ISI) in our initial experiments. This interval aimed to align with the internal alpha frequency, yet proved too short for reliable EEG analysis. Consequently, in our second experiment, we extended the ISI to 400 ms. However, this adjustment introduced potential trade-offs (Frensch et al., 1998; Frensch & Miner, 1994; Willingham et al., 1997), as a longer timeframe might impact participants' performance in our challenging task with rapid event occurrences that required the learning to be pure perceptual and implicit.

In our paradigm along with all the analysis, we took into account the potential time frame suggested for predictive learning. Stimuli in the sequence presented synchronized with alpha frequency (100 ms), which has been suggested that this may entrain the internal alpha rhythm with the stimuli (Calderone et al., 2014; Mathewson et al., 2009; Nobre et al., 2007). In addition to maintaining consistent timing in our experimental design, we aligned our

analysis with the hypothetical timeline associated with predictive learning. More specifically, our cross-temporal results showed higher decoding accuracy starting 200 ms (equivalent to two alpha cycles) before the onset in the selected time window. Consistently, in RSA, we observed changes within a 200 ms time window that revealed an enhancement of relevant information for the upcoming interval and a decrease in similarity with the other less relevant source of information, but still with a negative correlation suggesting a potential interplay between the two. Future research can take a look into this interplay by utilizing methods with more spatial resolution to explore how such information is stored and interacts maybe even in orthogonal subspaces (Panichello & Buschman, 2021). Overall, even with this consideration about the timing in our paradigm, precisely assessing the timing of this 100 ms cycle requires further investigation, making it an interesting question for future research.

An additional challenge in EEG processing pertains to the selection of a baseline, which can impact the final results (Alday, 2019; Widmann et al., 2015; Tanner et al., 2015, 2016; Zhang et al., 2021). Conventionally, researchers often prefer to use the time period preceding stimulus presentation as the baseline (Luck, 2022; Zhang et al., 2021). In our specific case, the logical choice would have been the 400 ms blank pre-stimulus time window [-400:0] ms. However, this time window coincides with the period of our interest. Consequently, we chose not to use this time window as our baseline, allowing us to capture the information processing dynamics during this period, rather than treating it as a reference point for comparison. Even though we finally used the time frame during which the sensory information was on the screen, it is consistent with the baseline-selection assumption that there are no systematic differences between conditions in the baseline interval (Widmann et al., 2015; Tanner et al., 2015, 2016; Luck, 2022; Zhang et al., 2021). Additionally, we performed additional analysis (no-baseline and baseline [-700:-500] ms) to ensure the reliability of our results with [0:100] ms baseline correction. The results from both options showed consistently high decoding accuracy, confirming that the selection of our baseline was not the primary factor influencing our decoding results.

In the RSA analysis, it is worth noting that the hypothetical RSA for previous locations was identical to the next hypothetical RSA. This could potentially create a challenge in determining whether the information similarity pertains to the past or the future. However, this challenge is mitigated when employing the cross-temporal analysis. In this approach, the model is trained on the pre-stimulus time window and then tested on the post-stimulus time window, which establishes a more cohesive link between the representation and information of the previous and next trials.

Furthermore, it is important to clarify that the ability to distinctly separate information related to the previous and next location is not an essential prerequisite to linking these findings to predictive learning. However, designing analyses to more thoroughly distinguish these two sources of information could be an interesting avenue for future research. Therefore, while each of these analyses in isolation may not be sufficient, collectively, they offer evidence supporting the involvement of predictive learning.

Taken together, our study employs a comprehensive combination of experimental and computational approaches to explore mechanisms associated with predictive learning in the brain. This research contributes to expanding our understanding of this crucial learning process, which has been suggested to be instrumental in the development of high-level abstract representations, which in turn enable the process of generalization to novel tasks.

2.6 Acknowledgments

We would like to thank all of the members of the Computational Cognitive Neuroscience Lab, Luck Lab at UC Davis, and Curran Lab at CU Boulder for their valuable and ongoing discussions on the topics explored in this work. This work was funded by ONR N00014-14-1-0670 / N00014-16-1-2128, ONR N00014-19-1-2684 / N00014-18-1-2116, N00014-20-1-2578, and ONR N00014-18-C-2067.

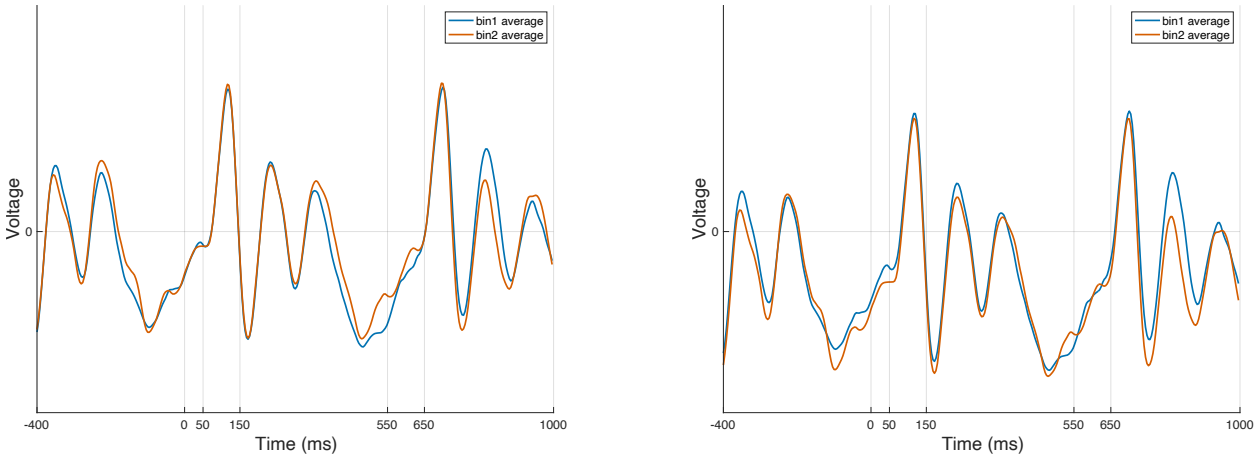
2.7 Supplementary materials

ERP results The following section presents the results of our event-related potential (ERP) analysis. Our study focused on two conditions: one with a baseline correction applied, where the baseline period was set from 0 to 100 ms (Figure 2.13a), and another condition without baseline correction (Figure 2.13b). Furthermore, we provide results for each group separately, comparing the effects of baseline correction on the recorded data in Figure 2.13.

In general, applying baseline correction from 0 to 100 ms led to more reliable and interpretable ERP results. This correction procedure effectively reduced unwanted fluctuations, allowing for a focused analysis of the neural activity specifically related to our study.

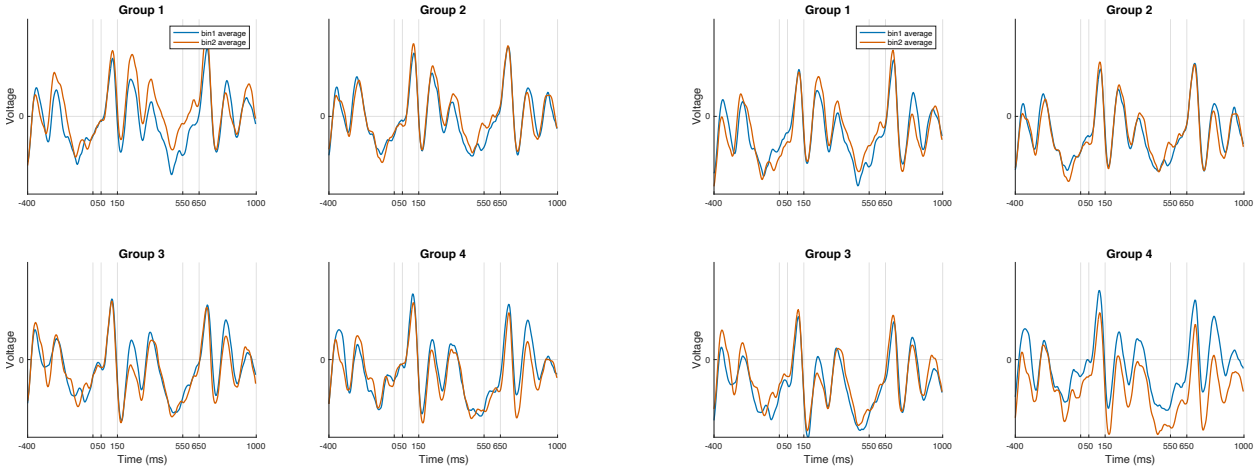
To select the most suitable baseline option, we considered factors such as proximity to the stimulus onset and the avoidance of confounding information. We evaluated alternative baseline periods of [-400:0] and [100:500] ms. However, the [-400:0] ms option included confounding information from the previous locations, which was not desirable. Additionally, the [100:500] ms period occurred before the next locations were presented and potentially contained information related to predictive learning, which was the focus of our study. As a result, we opted for the [0:100] ms baseline period based on these considerations. Although the [0:100] ms baseline period included the current stimulus on the screen, it provided the most appropriate baseline given that both bin1 and bin2 shared the same source of information for each location (e.g., loc 1 in group 1 for both bin 1 and bin 2). Therefore, considering the aforementioned reasoning, the [0:100] ms baseline period was the best option.

The inclusion of baseline correction significantly improved the quality of our ERP analysis. By minimizing confounding factors, this procedure enhanced the validity of our findings. Therefore, our analysis is based on the results obtained with the baseline correction applied.



(a) **(b)**

Figure 2.13: Comparison of ERP signals with and without baseline correction. (a) Presents the ERP signals with baseline correction [0:100] ms, and (b) Illustrates the ERP signals without any baseline correction.



(a) **(b)**

Figure 2.14: Comparison of ERP signals with and without baseline correction across each group. (a) Shows the ERP signals with baseline correction [0:100] ms. (b) Includes the ERP signals without any baseline correction.

ERP-based decoding results In addition to the analysis for all the groups with baseline 0:100ms, separate analyses were conducted for each group which further strengthened our observations (Figure 2.15a). This result provided compelling evidence of above-chance location decoding across most of the epoch period for ERP-based decoding across four groups

for 24 subjects and all used one-sample t-tests (group 1 : $t(23) = 6.0021, p = 0.00368$; group 2 : $t(23) = 5.2500, p = 0.00604$; group 3 : $t(23) = 4.9151, p = 0.00984$; group 4 : $t(23) = 4.0777, p = 0.01174$). Therefore, consistent with our previous analyses, the location of the target could be decoded across all four groups, as indicated by sustained ERP-based decoding accuracy. Similarly, when we did not have any baseline correction, the result for each group reached a significant above-chance decoding accuracy as illustrated in Figure 2.15b (group 1 : $t(23) = 5.0275, p = 0.00756$; group 2 : $t(23) = 5.4937, p = 0.00475$; group 3 : $t(23) = 4.3036, p = 0.00787$; group 4 : $t(23) = 4.3087, p = 0.00937$). It is important to note that the decoding accuracies across different groups exhibited slightly lower accuracy compared to the aggregated decoding accuracies over all groups, which is due to losing some power.

Further investigations into separated locations (Figure 2.15a) provided compelling evidence of above-chance location decoding across most of the epoch period for ERP-based decoding across all four groups for 24 subjects (group 1 : $t(23) = 6.0021, p = 0.00368$; group 2 : $t(23) = 5.2500, p = 0.00604$; group 3 : $t(23) = 4.9151, p = 0.00984$; group 4 : $t(23) = 4.0777, p = 0.01174$). Therefore, consistent with our previous analyses, the location of the target could be decoded across all four groups, as indicated by sustained ERP-based decoding accuracy. Similarly, when we did not have any baseline correction the result for each group reached significant above-chance decoding accuracy as illustrated in Figure 2.15b (group 1 : $t(23) = 5.0275, p = 0.00756$; group 2 : $t(23) = 5.4937, p = 0.00475$; group 3 : $t(23) = 4.3036, p = 0.00787$; group 4 : $t(23) = 4.3087, p = 0.00937$).

It is important to note that the decoding accuracies across different groups exhibited slightly lower accuracy compared to the aggregated decoding accuracies over all groups which is due to losing some power.

Decoding results for different baseline options across each group The result is presented in Figure 2.15 to show our findings with and without baseline correction for each of the four

groups.

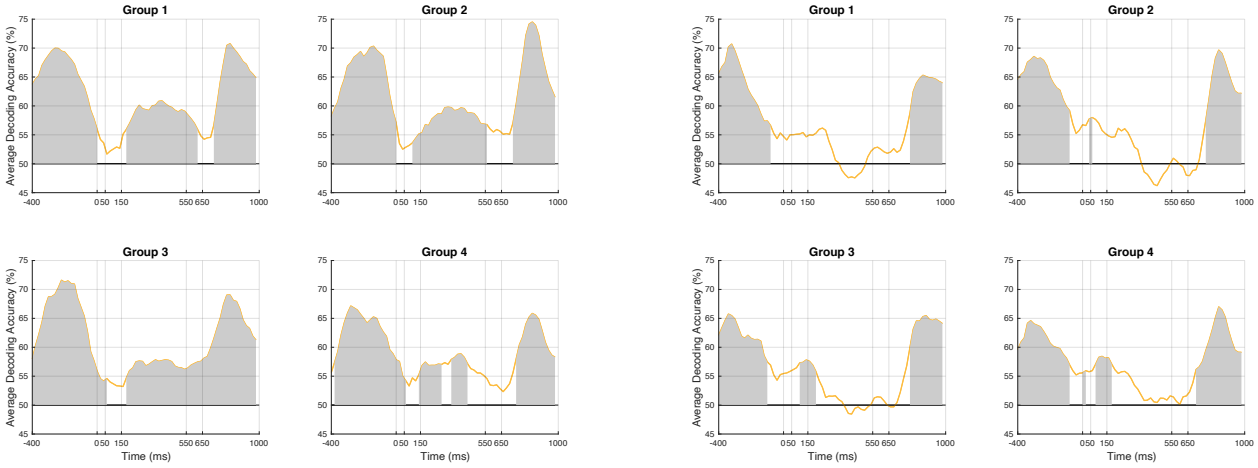
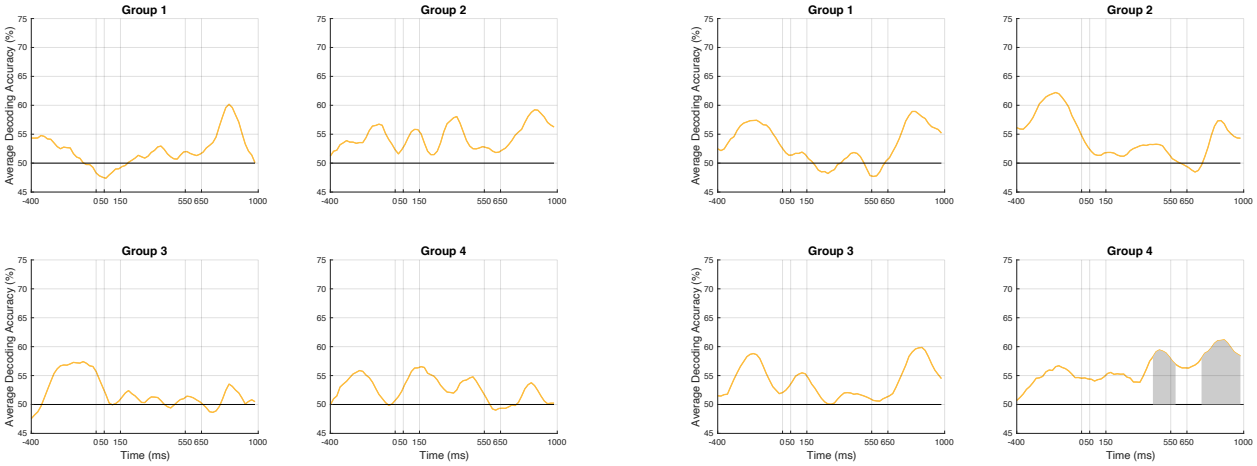


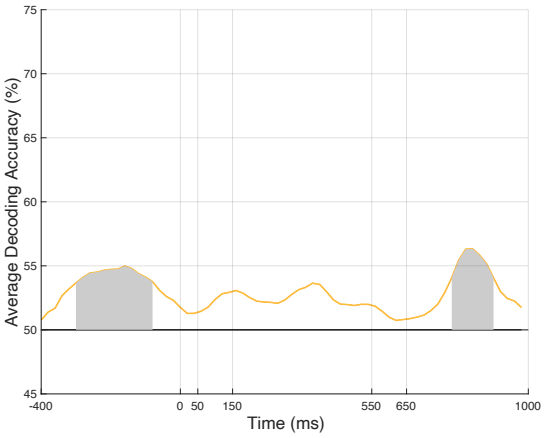
Figure 2.15: Comparison of ERP decoding accuracies with and without baseline correction across each group. (a) Shows the ERP decoding accuracies with baseline [0:100] ms. (b) Presents the ERP decoding accuracies without baseline correction.

Decoding results for different frequency bands The results are presented in Figure 2.16 and Figure 2.17.

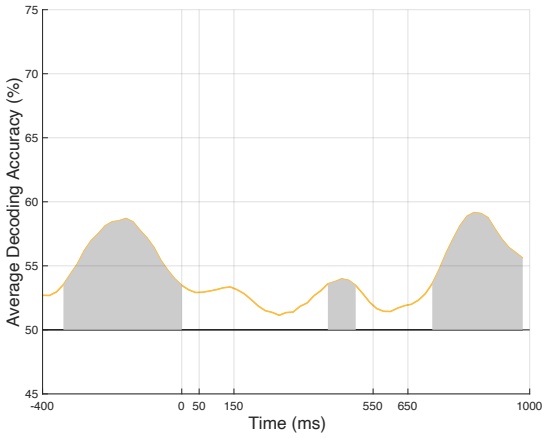


(a)

(b)

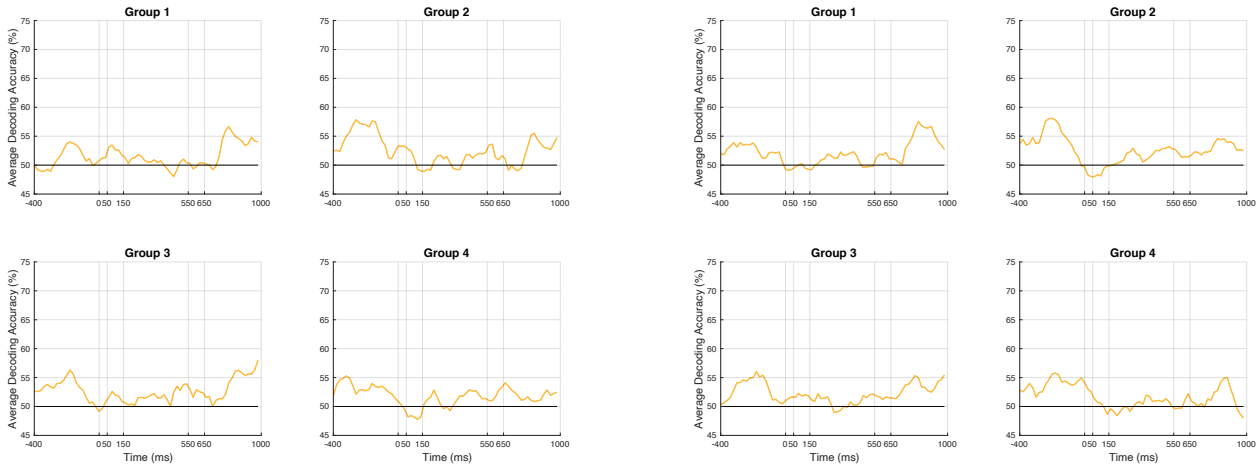


(c)



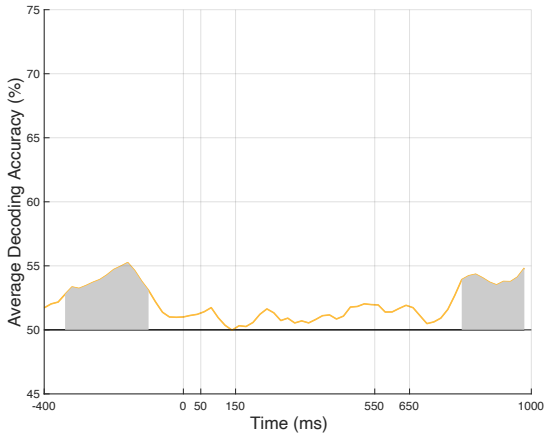
(d)

Figure 2.16: Decoding accuracies for theta and alpha frequencies with baseline correction [0:100] ms. Panels (a) and (b) show decoding accuracies of theta and alpha frequencies across four groups, respectively. Panels (c) and (d) present the aggregated decoding accuracies for theta and alpha frequencies, respectively.

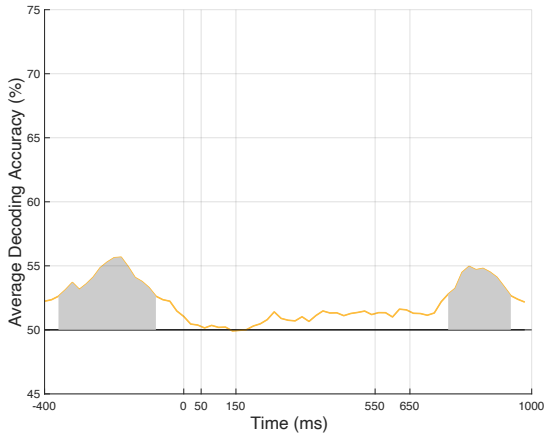


(a)

(b)



(c)



(d)

Figure 2.17: Decoding accuracies for beta and gamma frequencies with baseline correction [0:100] ms. Panels (a) and (b) display decoding accuracies across four groups for beta and gamma bands, respectively. Panels (c) and (d) present the aggregated decoding accuracies over four groups for beta and gamma frequencies, respectively.

Topographic plots Topographic plots for signals with baseline [0:100] ms are presented in Figure 2.18 and the topographic plots for signals without baseline correction are illustrated in Figure 2.19.

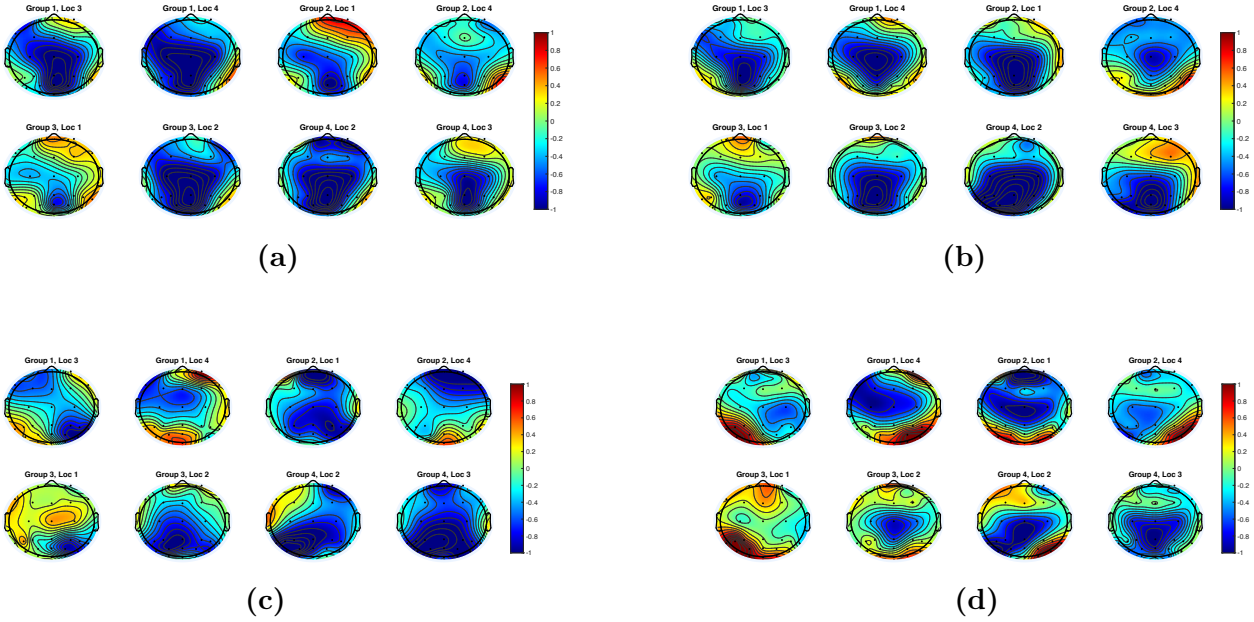


Figure 2.18: Topographic plots for signals with baseline correction [0:100] ms (the topographic maps for the time interval [0:100] ms have not been included since the activities during the baseline are close to zero). The plots display topographic maps for four groups, each containing two bins. For instance, the two top left topographic maps represent group 1, with location 3 corresponding to bin 1 and location 4 corresponding to bin 2. Subfigure (a) shows the topographic maps for the time interval [-400:0] ms, subfigure (b) displays the topographic maps for the time interval [150:550] ms, subfigure (c) exhibits the topographic maps for the time interval [550:650] ms, and subfigure (d) presents the topographic maps for the time interval [650:996] ms.

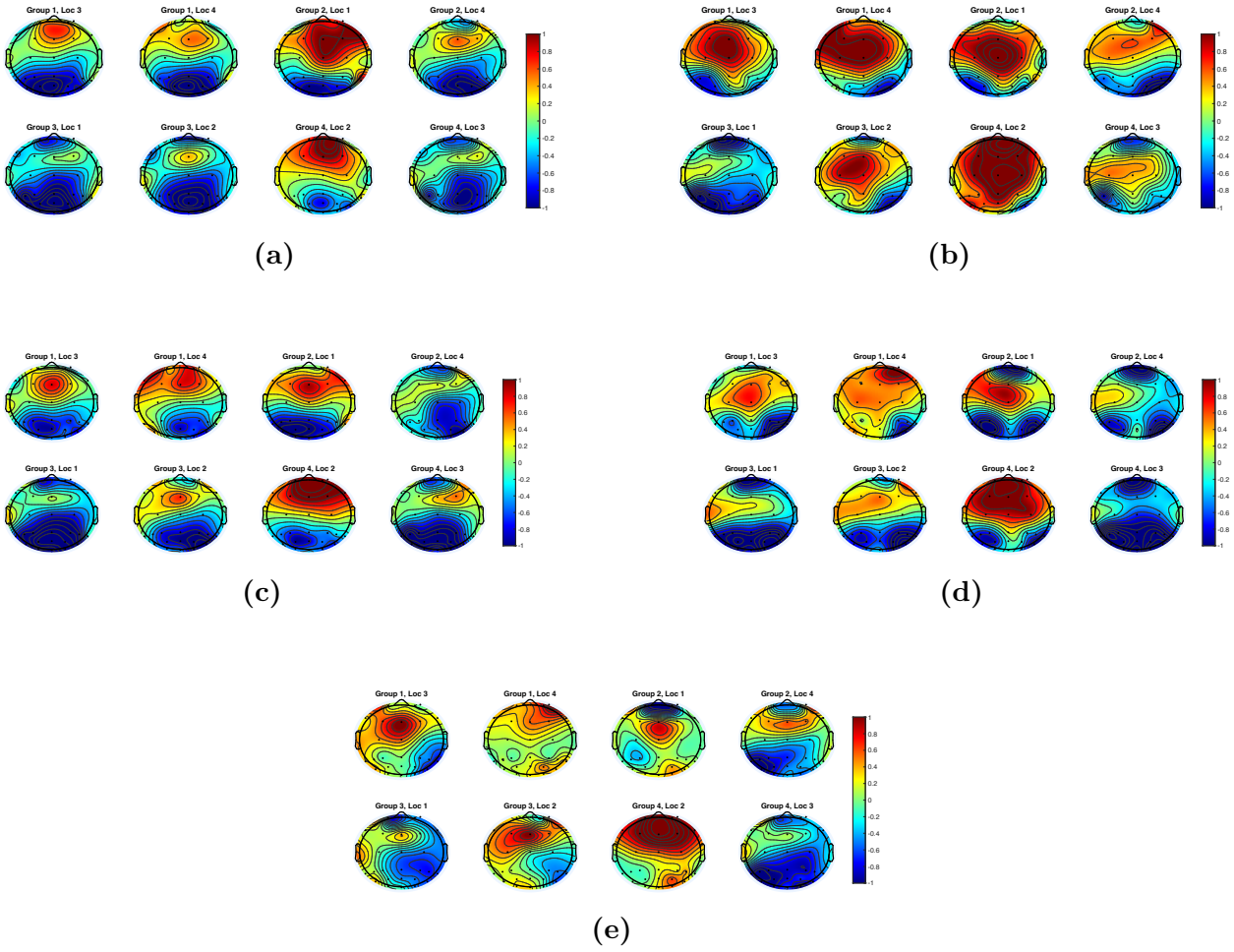


Figure 2.19: Topographic plots for signals without baseline correction. The plots display topographic maps for five time intervals, considering the 50 ms shift required for the information to reach the cortex: (a) [-400:0] ms, (b) [0:100] ms, (c) [150:550] ms, (d) [550:650] ms, and (e) [650:996] ms. The two top left topographic maps represent group 1, with location 3 corresponding to bin 1 and location 4 corresponding to bin 2.

Chapter 3

The Geometry of Map-Like Representations under Dynamic Cognitive Control[1]

Maryam Zolfaghar*[2], Jacob Russin*[2], Seongmin A. Park*[3], Erie Boorman[3], Randall C. O'Reilly[2] (* denotes equal contribution)

3.1 Abstract

Recent work has shown that the brain organizes abstract, non-spatial relationships between entities into map-like representations. However, an animal's objectives often depend on only a subset of the features of the environment. Under these circumstances, cognitive control – the capacity to flexibly select the features most relevant in the current context – becomes paramount. Here, we explore the relationship between cognitive control and the geometry of map-like representations by combining fMRI with neural network modeling. We find that brain areas including hippocampus and entorhinal cortex spontaneously organize pairwise relationships into 2D map-like representations, and that this 2D structure was controlled by compressing task-irrelevant dimensions in areas of prefrontal and parietal cortex. Our neural network model reproduced these findings and additionally predicted warping in the geometry along a context-invariant axis. This prediction was confirmed with fMRI, which

[1*] This chapter was originally accepted for publication in the Proceedings of the 44th Annual Meeting of the Cognitive Science Society (CogSci 2022) (Zolfaghar et al., 2022). The opinions expressed here are solely those of the author and do not necessarily reflect the official views of the conference, workshop, or publisher. The original version can be accessed online at: `https://escholarship.org/uc/item/28j425kf`.

[2] Center for Neuroscience, University of California, Davis

[3] Center for Mind and Brain, University of California, Davis

showed that the degree of warping was correlated with individual differences in cognitive control.

3.2 Introduction

Substantial evidence suggests the brain organizes incoming relational information into cognitive maps (O'Keefe & Nadel, 1978; Moser et al., 2008) – even when relations are abstract or non-spatial (Behrens et al., 2018; Bernardi et al., 2020; Garvert et al., 2017; Knudsen & Wallis, 2021; Park et al., 2020; Stachenfeld et al., 2017). For example, when humans are trained on stimulus features that vary systematically (e.g., the lengths of a bird's neck or the competence and popularity of people in a social hierarchy), brain areas such as the medial temporal lobe and medial parietal cortex efficiently encode these features with map-like representations (Constantinescu et al., 2021; Park et al., 2021, 2020). Just as geographic maps depict true distances between locations in the world, the geometry of these "map-like" representations reflects the latent structure of their underlying feature spaces, such that distances in the representational space are consistent with distances in the feature space. This map-like quality is thought to facilitate relational reasoning behavior, allowing animals to generalize to novel or unseen stimuli (Behrens et al., 2018; O'Reilly et al., 2021a; Summerfield et al., 2020; Whittington et al., 2020).

Most models of cognitive map formation consider cases where animals must navigate or reason within a single context or goal. However, animals are often faced with scenarios where multiple possible objectives can determine the subset of stimulus features that are important at any given time. These scenarios require cognitive control, or the capacity to flexibly select or attend the features of the environment that are most relevant to the current context or goal. Classic theories of cognitive control hypothesize that top-down attention mechanisms in the prefrontal cortex can dynamically modulate the processing in posterior brain regions in order to meet the demands of a current goal (Miller & Cohen, 2001; Herd

et al., 2006; Rougier et al., 2005). Computational models of these processes have been used to successfully explain cognitive and neural phenomena and have emphasized their functional benefits for managing the interference caused by conflict or incongruence (Miller & Cohen, 2001; Musslick et al., 2017; Shenhav et al., 2013). These studies have focused on classic cognitive control tasks such as the Stroop task (Stroop, 1935) that use discrete exogenous stimulus features such as color or orthographic features (Cohen et al., 1990; Herd et al., 2006; Rougier et al., 2005). However, less is known about how such processes might be used to control endogenous map-like representations retrieved from memory, such as those observed in the medial temporal lobe (MTL) or parietal cortex during abstract relational reasoning tasks.

Here, we combined fMRI with a neural network model to investigate the relationship between cognitive control and endogenous map-like representations. We tested human participants and the neural network on the same task, which was designed to facilitate the learning of map-like representations while simultaneously requiring the use of cognitive control as a function of the current task context. Using parallel analyses of the neural network and human fMRI data, we observed three key phenomena related to cognitive maps, cognitive control, and their relationship:

1. Learning in both the human brain and neural network models sculpted unitary map-like representations that integrated information across multiple contexts to capture the latent relational structure of the task space.

2. To effectively resolve the interference caused by incongruence in the task, these map-like representations were dynamically modulated by cognitive control processes such that irrelevant dimensions of the representational space were compressed according to the current context.

3. This interference, and the resultant demand on cognitive control, was related to congruence effects in the map-like representations learned over the course of the task, as

measured by the degree of warping in their geometry: pattern dissimilarity between congruent stimulus pairs was greater than that of incongruent stimulus pairs.

3.3 Methods

3.3.1 Experimental Task

Participants learned about the relative ranks of 16 hypothetical people on two social hierarchy dimensions: "competence" and "popularity." Unknown to the participants, these 16 people were arranged in a 4x4 grid along these two axes (see Figure 3.1). On each trial, a cue indicating the axis was presented, followed by two images showing the faces of two of the people in the hierarchy. The stimuli consisted of 16 grayscale photographic images of faces (Strohminger et al., 2016) and two colored cues (red and blue squares). Participants were instructed to select which of the two people ranked higher on the indicated axis. During training, participants only saw pairs of faces that differed by one rank along the appropriate axis (and were instructed that this was the case). Participants were then tested in the scanner on pairs with rank differences greater than or equal to one, requiring them to make novel transitive inferences.

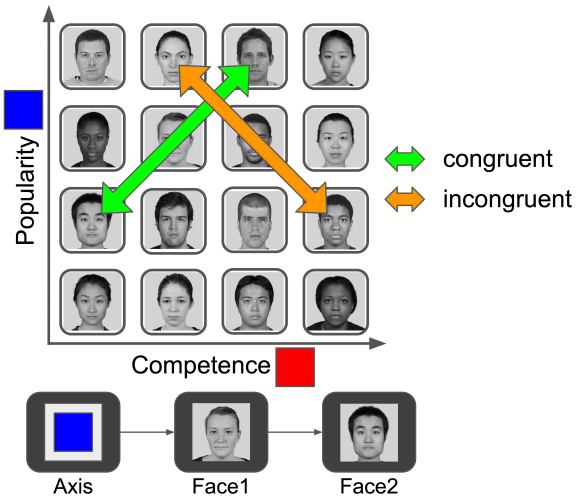


Figure 3.1: Experimental task design. Participants made decisions about which of two people ranked higher on one of two social hierarchy dimensions. The dimension ("Axis") was cued at the start of each trial. Unknown to the participants, the 16 hypothetical people in the hierarchy were arranged in a 4x4 grid along these two dimensions. Some pairs were congruent (example shown with green arrow), in the sense that the correct face ranked higher on both dimensions, and some pairs were incongruent (example shown with orange arrow), in the sense that each of the two faces ranked higher on one of the two dimensions, thereby requiring the "Axis" cue to disambiguate the higher-ranking face.

It was hypothesized that over the course of training, participants would learn the latent 4x4 grid structure of the hierarchy. However, participants could only learn this structure through pairwise comparisons, allowing us to investigate how the brain learns to encode stimuli into map-like representations.

In addition to facilitating an examination of cognitive maps, the task probed the interaction between these map-like representations and cognitive control processes. Cognitive control is required to selectively attend to goal-relevant features in the presence of interference (Miller & Cohen, 2001). In this task, the relevant axis of the social hierarchy is cued, but the irrelevant axis may cause interference when the pair of faces is incongruent, i.e., when each of the two faces being compared ranks higher on one of the two axes, such that the correct answer depends on which axis was cued (see Figure 3.1). This is analogous to classic cognitive control tasks such as Stroop (Stroop, 1935), where control demands are

increased when the stimulus features (e.g., ink color and color word) conflict.

3.3.2 Participants

A total of 33 participants (16 female, age range: 19–23, normal or corrected to normal vision) were recruited for this study via an online recruitment system. Six participants were excluded due to strong head movements larger than the voxel size of 3mm. In total, 27 participants entered the analysis (mean age: 19.37 $\pm$ 0.26, standard error mean (SEM)). The study was approved by the local ethics committee, all relevant ethical regulations were followed, and participants gave written consent before the experiment.

3.4 Neural Network Model

The model was trained and tested on the same task used in the fMRI experiment, including its 4x4 grid structure and transitive inference test. On each trial, the model was presented with two 64x64 grayscale images of faces (x_1 and x_2), along with the context indicating the axis along which they should be compared (x_a - represented as a 1-hot vector). The model was trained to select which of the two faces ranked higher on the appropriate axis through supervised feedback on the correct answers during training. As in the fMRI experiment, the model did not have access to these rankings or to the latent structure of the 4x4 grid and had to learn these through trial-and-error on pairwise comparisons.

Our goal was to explore the effects of dynamic cognitive control on the geometry of map-like representations, so we developed a recurrent neural network model in order to capture the dynamics of representations unfolding over the course of each trial.

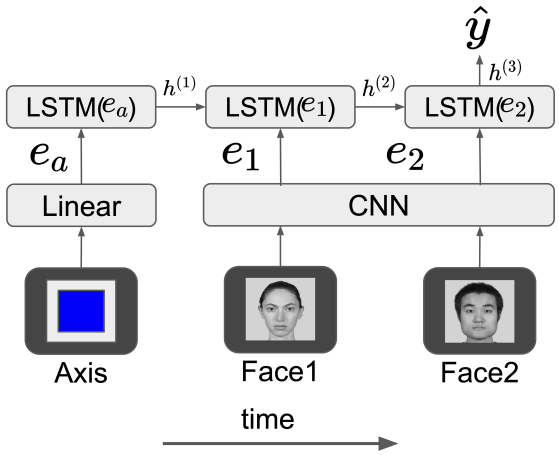


Figure 3.2: Neural network model architecture. The model used a convolutional neural network (CNN) to process images and a long short-term memory (LSTM) to process the sequence of inputs over time.

3.4.1 Model Architecture

The neural network model (see Figure 3.2) was composed of standard building blocks including a convolutional neural network (CNN) and a long short-term memory (LSTM). The CNN processed the images of faces, and a linear embedding layer processed the axis. The same CNN was used to process both faces on each trial. The result of these initial operations was a single vector encoding each of the axis (e_a), Face 1 (e_1), and Face 2 (e_2), each with the same dimension:

$$e_a = W_a x_a \quad e_1 = CNN(x_1) \quad e_2 = CNN(x_2) \tag{3.1}$$

where x_a is a 2-dimensional (one for each context) one-hot vector indicating the current context, and x_1 and x_2 are the Face 1 and Face 2 images. The LSTM processed the embeddings

(e_a, e_1, e_2) in sequence:

$$h^{(t)} = \text{LSTM}(e^{(t)}, h^{(t-1)}) \tag{3.2}$$

$$\hat{y} = W_o h^{(3)} \tag{3.3}$$

where $h^{(t)}$ is the hidden state of the LSTM at time-step t, $e^{(t)}$ is the input embedding at time step t (e_a, e_1, or e_2), and W_o is a linear output layer that produces a prediction $\hat{y}$ about the answer from the final (third) hidden state of the LSTM ($h^{(3)}$).

3.4.2 Implementation Details

All modeling experiments were implemented using PyTorch. The model was trained using standard optimization techniques, including the backpropagation algorithm. The model was optimized using Adam (Kingma & Ba, 2015) with a learning rate of 0.001 and a batch size of 32 for 1000 gradient steps. For each simulation, 20 runs were performed with different random initializations.

The CNN included two convolutional layers with kernel sizes (3, 3), strides (2, 2) and number of channels (4, 8). Max pooling followed each convolutional layer with kernel sizes (2, 2) and strides (2, 2). The CNN also included a final linear layer to map the output of the last pooling operation to a single vector with 32 dimensions. The axis embedding e_a was also 32-dimensional, and the LSTM had a hidden layer size of 128.

3.5 Results

Participants performed well on the unseen pairs of faces tested in the scanner (93.6% mean accuracy), indicating good transitive inference performance. To test our main hypotheses, we conducted representational similarity analyses (RSA) Kriegeskorte et al. (2008) to characterize the representations in the brain, and performed analogous tests on the representations

gathered from the model throughout training (see Figure 3.3). In the following, we describe the results of the analyses pertaining to each of our main hypotheses, with particular emphasis on the warped representational geometry, as this was the most novel aspect of our investigation.

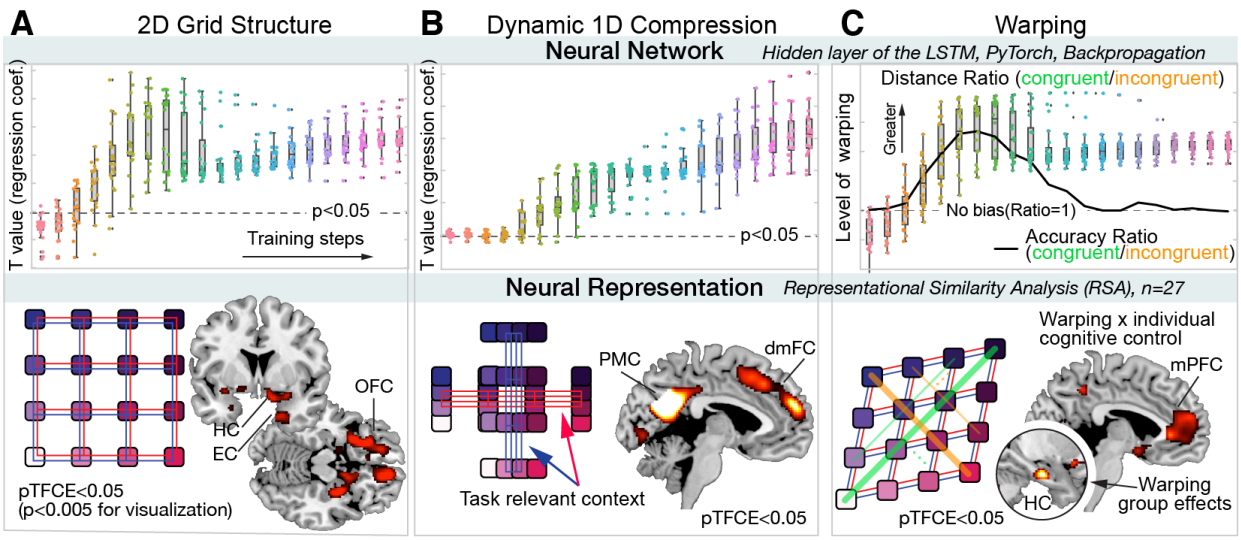


Figure 3.3: Results of analyses testing three main hypotheses. Top panels show results of analyses on the neural network model, which are given by statistics calculated on its representations over the course of training. Box plots show the variance in these statistics over 20 runs. Bottom panels show results of fMRI analyses, along with diagrams depicting idealized representations of the grid used as regressors in the RSA. **A)** Evidence of map-like representations was found in both the model and the brain. In the model, this is shown by a significant relationship ($p< 0.05$) between distances between representations and their corresponding distances in the underlying grid. In the fMRI analysis, map-like representations were found in hippocampus (HC), entorhinal cortex (EC) and orbitofrontal cortex (OFC). pTFCE: threshold-free cluster enhancement. **B)** Evidence of dynamic selection of the task-relevant dimension was found in both the model and the brain. Representations were expanded along the task-relevant axis (or equivalently, compressed along the task-irrelevant axis). This effect emerged early in the model's training and was significant in posterior and medial parietal cortex (PMC) and dorsomedial frontal cortex (dmFC). **C)** Evidence of warped representational geometry was observed in both the model and the brain. Distances between representations of congruent pairs of faces (along the green diagonal) were larger than those of incongruent pairs (along the orange diagonal), causing a consistent warping of the space along the context-invariant axis (i.e., along the congruent diagonal). In the model, this is visualized with the ratio of average Euclidean distances between congruent pairs of faces to average Euclidean distances between incongruent pairs of faces, which is consistently larger than 1 throughout training. This was associated with the corresponding accuracy ratio (black line in the same plot). Warping was observed on a group-level in HC, and warping in the medial prefrontal cortex (mPFC) and posterior cingulate cortex (PCC, not labeled in figure) was correlated with individual differences in cognitive control (as measured by differences in reaction times between congruent and incongruent trials.)

3.5.1 Map-Like Representations

Although neither the model nor the human participants were explicitly instructed on the underlying structure of the 4x4 grid, representations in hippocampus (HC), entorhinal cortex (EC), orbitofrontal cortex (OFC), and in hidden layers of the model captured this basic structure (see Figure 3.3A). In both the model and these brain regions, similarity between representations was correlated with 2D Euclidean distances between faces' position in the 4x4 social hierarchy. In the fMRI data, this was shown with both whole-brain searchlight-based and anatomically-defined ROI-based RSA. Importantly, ROI analyses revealed that idealized 2D representations explained pattern similarity significantly better than the alternative hypothesis of two separate one-dimensional maps in HC, EC, and OFC. In the model, a similar analysis revealed a significant correlation between representational distances and pairwise Euclidean distances between faces in the 4x4 grid ($p < 0.05$). This relationship emerged early and was maintained throughout training. These results are consistent with the hypothesis that the human brain spontaneously organizes incoming relational information into map-like representations (Behrens et al., 2018; Park et al., 2020), and show that these emerge in a neural network model without any specialized components that were explicitly designed to do so.

3.5.2 Dynamic Selection of Task-Relevant Dimension

This basic 2D representational geometry was dynamically modified by the current context: in both the neural network model and in brain regions including dorsomedial frontal cortex (dmFC) and posterior and medial parietal cortex (PMC) distances along the irrelevant axis were compressed (see Figure 3.3B). In the fMRI data, this was again shown using a searchlight-based multiple regression RSA that included the representational dissimilarity matrix (RDM) capturing 2D Euclidean distances between face pairs and a RDM capturing task-relevant 1D distances that assumed the currently irrelevant dimension of the grid was compressed (see Figures 3.3A and 3.3B, respectively). In the model, a regression re-

vealed a significant relationship between the pairwise 1D rank-differences between faces in the task-relevant dimension and the distances between corresponding representations (over and above their 2D structure, which was also included as an independent variable in the regression). These findings indicate that the task-irrelevant features of the map-like representations stored in memory were compressed relative to the task-relevant features. This relationship emerged early in training and was maintained throughout the entire training period. These findings are consistent with previous experiments in tasks with exogenous sensory features in these regions (Mante et al., 2013; Takagi et al., 2021; Flesch et al., 2022a), and suggests how cognitive control can operate on 2D map-like representations from memory by accentuating the task-relevant dimensions of a learned representational space.

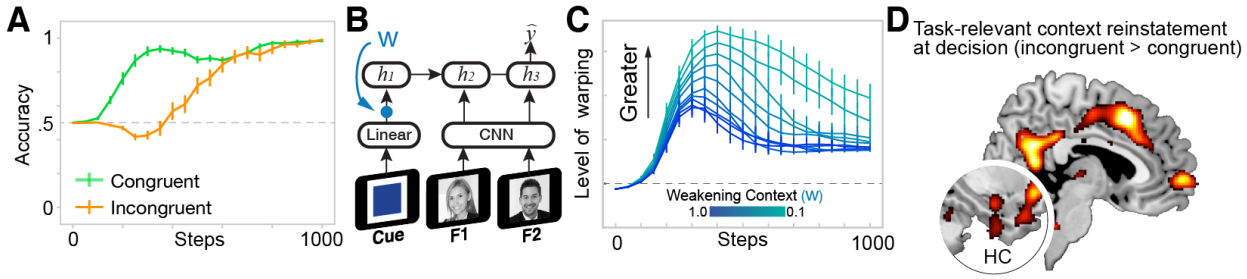


Figure 3.4: Results of additional analyses addressing the causal relationship between warping and cognitive control. **A)** The model improved its accuracy on congruent trials before incongruent trials. This difference coincided with the emergence of warping in its representations (see accuracy ratio vs. warping in previous figure), suggesting that warping is associated with a decreased reliance on contextual information and cognitive control. Error bars show standard error of the mean (SEM). The dotted horizontal line indicates chance performance (50% accuracy). **B)** To directly investigate the relationship between the strength of contextual information and the degree of warping, we performed simulated ablations of the axis embedding in the model by multiplying (e_a) by values (w) ranging from 0 to 1. **C)** Results of this ablation experiment showed increases in warping (again measured by the ratio of congruent distances to incongruent distances) when context information was inhibited. The dotted horizontal line indicates a ratio of 1 (i.e., no warping). Error bars again indicate SEM. **D)** A related fMRI analysis showed stronger reactivation of the current context on incongruent trials than congruent trials in hippocampus (HC), medial prefrontal cortex (mPFC) and posterior cingulate cortex (PCC) at decision time (pTFCE$<$ 0.05; p$<$ 0.005 for visualization).

3.5.3 Warped Representational Geometry

In the fMRI experiment, reaction time (RT) was faster on congruent than incongruent trials regardless of the distance between locations of faces ($p < 0.01$), providing a measure of individual differences in cognitive control. We also tested for an effect of congruence on the map-like representations in both the model and the human participants by comparing distances between representations of congruent and incongruent pairs of faces sampled from different trials across blocks. Analyses of both the hidden layers and fMRI data consistently showed warping along the congruent compared to the incongruent axis – i.e., a stronger relationship between pattern similarity and Euclidean distance in congruent pairs of faces than incongruent pairs of faces (see Figure 3.3C). In the model, a regression revealed that distances between representations of congruent pairs of faces were significantly larger than those of incongruent pairs ($p < 0.05$, see top panel of Figure 3.3C). Group-level fMRI analyses revealed the same warped geometry in representations in brain areas including HC (see bottom panel of Figure 3.3C). Additionally, the levels of warping observed in the amygdala, medial prefrontal cortex (mPFC), and posterior cingulate cortex (PCC) were correlated with individual differences in our behavioral measure of cognitive control (i.e., the difference in RT between congruent and incongruent trials). This result suggests that individuals with greater representational warping in these brain regions also experienced a greater demand on cognitive control on incongruent trials (as measured in their behavior).

Finally, we found that the warping in the neural network model was linked to an initial tendency to ineffectively utilize context information that would mitigate the interference on incongruent trials. On congruent trials, context information does not strictly need to be maintained because the correct answer does not depend on the current context (see Figure 3.1). However, on incongruent trials context information is required to mitigate the interference caused by the fact that each face ranks higher on one of the two axes. We observed that early in training, the model performed well on congruent trials but poorly on incongruent trials (see Figure 3.4A), and that this difference emerged simultaneously with the warping

in its representational geometry (see accuracy ratio in Figure 3.3C). We reasoned that the warping may be related to the model's capacity to manage the interference on incongruent trials by utilizing context (axis) information. We therefore simulated an "ablation" of the contextual input by down-scaling the embedding vector (e_a) by multiplying it by a factor less than 1 (see Figure 3.4B). Results of this simulated ablation showed that warping increased when context was inhibited (see Figure 3.4C), confirming the link between the strength of context information in the model and the degree of warping in its representations. Inspired by this observation in the model, an fMRI analysis found stronger reactivation of the current context on incongruent trials than congruent trials in HC, mPFC and PCC at decision time (see Figure 3.4D), suggesting greater reinstatement of the behavioral context during decisions in the presence of incongruence.

3.6 Discussion

It has been suggested that a key to the power of human intelligence is the capacity to integrate sensory information into cognitive maps or representations that capture the structure of the environment (Behrens et al., 2018; O'Reilly et al., 2021a). The map-like quality of these representations is thought to empower deliberative or model-based reasoning capabilities and to facilitate generalization to unseen stimuli (Behrens et al., 2018; Vikbladh et al., 2019). However, humans are not only capable of systematically encoding relevant structural information into map-like representations, but can flexibly deploy them to meet the demands of multiple contexts or goals (Miller & Cohen, 2001; Musslick et al., 2017). This flexibility is thought to emerge in part from cognitive control mechanisms in the prefrontal cortex (Herd et al., 2006; Miller & Cohen, 2001; Rougier et al., 2005), which may endow humans with the specialized circuitry necessary for systematic generalization (Russin et al., 2020).

In this work we explore the relationship between cognitive control and the geometry of map-like representations by integrating fMRI with neural network models (Flesch et al.,

2018, 2022a). Consistent with previous work (Russin et al., 2021), we find that when a neural network was trained on our task, it developed map-like representations that captured the latent 2D structure of the task, qualitatively reproducing phenomena observed with fMRI in HC, EC and OFC. These representations emerged despite the fact that only one of the two dimensions of the grid was cued at a time and both the model and the human participants learned from pairwise comparisons alone.

As with all neural network simulations, our model could not be supplied with the wealth of experience that human participants bring to laboratory tasks such as ours, and therefore may not capture the breadth of the processes that are likely involved in the formation of cognitive maps in humans. We expect that prior experience with 2-dimensional spaces allowed participants to leverage existing knowledge while they learned the task. However, we emphasize that the latent 4x4 configuration of the faces was completely arbitrary, and that despite its lack of prior experience, the model captured the map-like qualities observed in the brain's representations.

The emergence of map-like representations allowed us to explore their interaction with cognitive control. In particular, we investigated whether their geometry was dynamically modulated according to the current context. Parallel analyses revealed dynamic compression of the irrelevant axis (or equivalently, expansion of the relevant axis) in the representations of both the neural network and brain regions including PMC and dmFC, consistent with previous experimental findings using a different task (Flesch et al., 2022a). This phenomenon emerged in the dynamics of the model through learning: the model was not explicitly designed with a capacity to scale its representations or implement a specific mechanism for cognitive control. However, these emergent dynamics are consistent with previous neural network models of cognitive control, which implement a top-down attention mechanism to modulate specific stimulus features in posterior representations (Herd et al., 2006; Rougier et al., 2005). Flesch et al. (2022a) found that a similar compression emerged in neural networks in the "rich" regime, where they were initialized with small weights. We did not test

our models with different initialization schemes, but we expect that the defaults we used would put them in the "rich" regime. Future work will test the extent to which our results depend on the magnitudes of initial weights.

The model also revealed another way in which cognitive control can affect the geometry of map-like representations: the 2D structure of the representations in the model was warped along the context-invariant axis (i.e. the congruent diagonal, see Figure 3.3C). Again, this phenomenon occurred without any specific modification to the model, suggesting it is an emergent property of representations learned by neural networks trained on the task. This prediction of the model was confirmed in the fMRI experiment: group-level analyses revealed significant warping in HC, and warping in mPFC and PCC was found to be correlated with individual differences in cognitive control.

One of the strengths of pairing a computational model with any experimental approach is the ability to simulate experiments that would be difficult or impossible with real subjects. Our simulations offered insight into the relationship between warping and the dynamics of cognitive control. Early in training, the model performed better on congruent than incongruent trials; because congruent trials did not require contextual information to be used, we reasoned that warping in the model's 2D map-like representations was related to its capacity to utilize the current context. When we simulated an ablation of contextual information, warping increased and was maintained for a longer time period throughout training (see Figure 3.4). This suggests that warping may be a natural way for the brain to compensate for a relatively weak capacity for cognitive control to orthogonalize representations according to the current context. Without strong contextual information, learning may opportunistically seize on relations between congruent pairs, which do not require context to disambiguate correct responses.

If no contextual information was available whatsoever, the best an agent could do would be to collapse the 2D grid into an integrated 1D ranking, resulting in a perfectly "warped" map projecting each face onto the congruent (bottom-left to top-right) diagonal. Thus, one

explanation for our findings is that warping in HC, as well as mPFC and PCC compensated for imperfect cognitive control in the human participants by shifting their representations to approximate this idealized 1D ranking. This is consistent with the finding that the degree of warping found in mPFC and PCC was correlated with individual differences in cognitive control, as measured by the difference in RT between congruent and incongruent trials. This explanation is also consistent with a further fMRI analysis that found that the current task-relevant context was more strongly reinstated in HC, mPFC and PCC on incongruent trials compared to congruent trials. An alternative way of interpreting our results is that participants who developed more warping in their representational geometry could not utilize context information as effectively during learning, which in turn led to greater difficulty in overcoming interference on incongruent trials. Our results, although they are suggestive, cannot definitively establish the direction of causality between representational geometry and individual differences in cognitive control. We leave it to future work to more thoroughly investigate the causal structure of the link between these phenomena that we establish here.

Taken together, our results reveal an intricate relationship between cognitive control during cognitive map formation, the resulting representational geometry, and its role on subsequent control during decisions. We found evidence of complementary representational geometries for efficiently encoding abstract relational information and flexibly selecting behaviorally relevant attributes from those representations in both neural networks and human brains. The findings further cast cognitive control in a new light, whereby an individual's representational geometry is both sculpted by and used for cognitive control when retrieving representations with endogenous feature dimensions from memory. Furthermore, our work demonstrates the virtues of integrating a neural network modeling approach with neuroimaging, and may help to address current limitations of modern neural networks used for artificial intelligence (Russin et al., 2020).

3.7 Acknowledgments

We would like to thank the members of the Computational Cognitive Neuroscience lab and the Learning and Decision Making lab, as well as reviewers for helpful comments and discussions. The work was supported by: ONR grants ONR N00014-20-1-2578, N00014-19-1-2684 / N00014-18-1-2116, N00014-18-C-2067, as well as NSF CAREER Award 1846578, and NIH R56 MH119116.

Chapter 4

A Neural Network Model of Continual Learning with Cognitive Control[1]

Jacob Russin[2], Maryam Zolfaghar[2], Seongmin A. Park[3], Erie Boorman[3], Randall C. O'Reilly[2]

4.1 Abstract

Neural networks struggle in continual learning settings from catastrophic forgetting: when trials are blocked, new learning can overwrite the learning from previous blocks. Humans learn effectively in these settings, in some cases even showing an advantage of blocking, suggesting the brain contains mechanisms to overcome this problem. Here, we build on previous work and show that neural networks equipped with a mechanism for cognitive control do not exhibit catastrophic forgetting when trials are blocked. We further show an advantage of blocking over interleaving when there is a bias for active maintenance in the control signal, implying a tradeoff between maintenance and the strength of control. Analyses of map-like representations learned by the networks provided additional insights into these mechanisms. Our work highlights the potential of cognitive control to aid continual learning in neural networks, and offers an explanation for the advantage of blocking that has been observed in humans.

[1*] This chapter was originally accepted for publication in the Proceedings of the 44th Annual Meeting of the Cognitive Science Society (CogSci 2022) (Russin et al., 2022). The opinions expressed here are solely those of the author and do not necessarily reflect the official views of the conference, workshop, or publisher. The original version can be accessed online at: `https://escholarship.org/uc/item/3gn3w58z`.

[2] Center for Neuroscience, University of California, Davis

[3] Center for Mind and Brain, University of California, Davis

4.2 Introduction

Neural networks have shown impressive performance in many domains in machine learning (ML), where they are typically trained on batches of data that are independent and identically distributed (Hadsell et al., 2020). However, agents learning about the world in real time experience streams of data that are not independent (e.g., a human may spend a few hours exploring one part of an unfamiliar city). The neural networks that have driven recent success in artificial intelligence perform poorly in these continual-learning settings because of the well known phenomenon of catastrophic forgetting/interference (McClelland et al., 1995; McCloskey & Cohen, 1989). When samples or trials are blocked, learning in new blocks overwrites the learning that occurred in previous blocks. Humans and other animals do not exhibit such extreme forgetting (McClelland et al., 1995), and in some cases even demonstrate an *advantage* when trials are blocked (Carvalho & Goldstone, 2014; Flesch et al., 2018; Noh et al., 2016; Wulf & Shea, 2002), suggesting there are mechanisms in the brain that mitigate catastrophic forgetting and can even reverse it, making learning easier when experiences are correlated over time.

A number of strategies for overcoming catastrophic forgetting in neural networks have been proposed in both computational neuroscience (Flesch et al., 2018, 2022b; McClelland et al., 1995) and ML (Botvinick et al., 2019; Hadsell et al., 2020; Mnih et al., 2013; Velez & Clune, 2017). Complementary learning systems (CLS) theory emphasizes that catastrophic forgetting arises when learning occurs too quickly in overlapping representations (McClelland et al., 1995; O'Reilly et al., 2011), and that the episodic memory system in the hippocampus plays an important role in learning representations that are sparse or pattern-separated, allowing rapid learning to take place. However, constraining patterns of activity to be sparse is not the only way to ensure they will not overlap and interfere with each other. Theories of cognitive control in the prefrontal cortex (PFC) emphasize that a crucial function of control is to selectively modulate activity in other brain areas in order to coordinate a response that aligns with the current context or goal (Herd et al., 2014; Miller & Cohen, 2001; Rougier

et al., 2005). Cognitive control may therefore play an important role in regulating learning so that patterns of activity do not overlap across different contexts or goals (Rougier et al., 2005; Tsuda et al., 2020).

Here, we build on this work and test neural networks in conditions where trials are either blocked or interleaved, showing how even in the absence of a hippocampal episodic memory system, cognitive control can help to mitigate catastrophic forgetting in the blocked condition. We further hypothesized that in some cases learning across blocked trials is *superior* to interleaving because of an internal bias of the PFC to maintain its activity over time, creating a cost to rapidly switching between contexts or goals (Blackwell et al., 2014; Herd et al., 2014; O'Reilly & Frank, 2006). This idea fits well with a general framework where the cost of switching must be traded off against the strength of control: stronger control results in less catastrophic forgetting, but more difficulty switching (Herd et al., 2006; Shenhav et al., 2013). We perform our simulations on a task designed to induce learning of map-like representations (Park et al., 2021, 2020; Russin et al., 2021) so that we could additionally investigate how cognitive control affected the model's representations.

4.2.1 Task

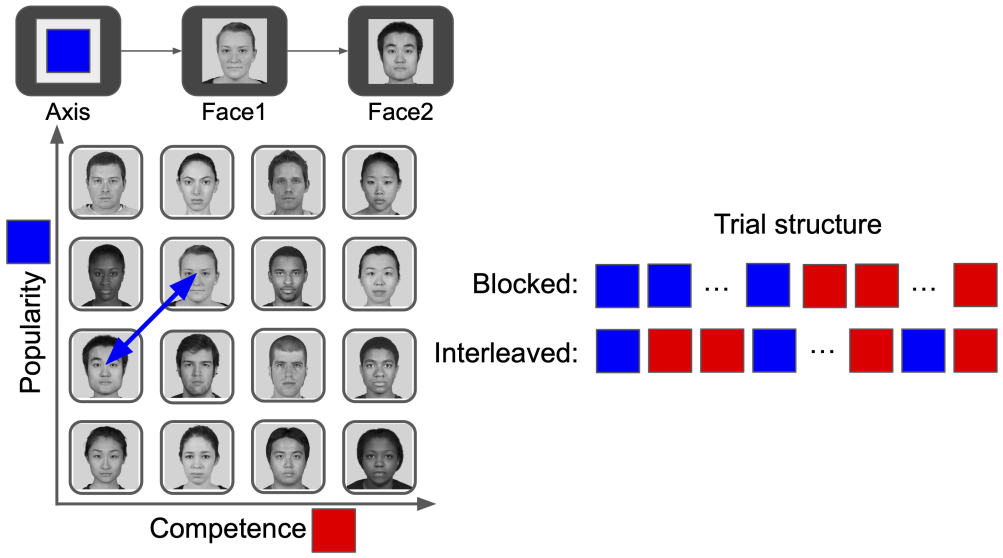


Figure 4.1: Task structure. The model learned the relative ranks of people along two social hierarchy dimensions: popularity and competence. The model learned through trial and error to select which of two faces ranked higher along one of the two dimensions (indicated by a cue). Trials were either interleaved, where cues were randomly shuffled, or blocked, where one dimension was learned at a time.

We trained neural network models on an existing task taken from an fMRI experiment (Park et al., 2020; Russin et al., 2021). Participants in the experiment learned about the relative ranks of 16 people in a hypothetical social hierarchy along two separate social dimensions: "popularity" and "competence" (see Figure 4.1). On each trial, the participants predicted which of two people ranked higher on one of the two dimensions, indicated by a cue. Unknown to the participants, these faces were organized into a 4x4 grid along the two dimensions; the participants were not instructed on the structure of the grid, and had to infer this structure from trial-and-error learning over pairwise comparisons.

During training, participants only saw pairs of faces that differed by one rank on the given dimension. Then in the scanner they performed a transitive inference test where comparisons were made between faces more than one rank apart. Intriguingly, the researchers found in pilot experiments that participants learned better when trials were blocked (i.e., one

dimension learned at a time). This is consistent with previous results showing that learning is improved when trials are blocked (Flesch et al., 2018). This task allowed us to explore the learning dynamics in our models, but because it was designed to investigate cognitive maps in the brain, we were also able to make concrete predictions about the representations that would be learned under different conditions.

We tested neural networks on the same task structure, including its 4x4 grid and transitive inference test. However, we introduced two training conditions to compare the learning behavior of models when trials were blocked vs. interleaved (see Figure 4.1). In the interleaved condition, popularity and competence trials were shuffled randomly, but in the blocked condition the models were trained on one of the two dimensions at a time. This allowed us to investigate the potential for cognitive control and gating mechanisms to alleviate the effects of catastrophic forgetting, as has been observed in humans learning certain tasks (Carvalho & Goldstone, 2014; Flesch et al., 2018; Noh et al., 2016; Wulf & Shea, 2002).

4.3 Neural Network Model

We designed a neural network that leveraged the principles of cognitive control in the PFC, including active maintenance and selective modulation according to the current context or goal. To test our hypotheses, we implemented models 1) with and without PFC gating, 2) with different levels of active maintenance, and 3) with different levels of control strength.

4.3.0.1 Base Model

To start, we built a simple base neural network with a multi-layer perceptron (MLP) for learning the relationships between the faces in the task (see Figure 4.2). The base model takes three one-hot vectors representing the context cue ("Axis", 2 dimensions) and each of the two faces ("Face1" and "Face2", 16 dimensions each) as inputs, and returns a prediction for which face ranked higher on the appropriate dimension. Each of these three inputs were

embedded with linear layers, concatenated, and fed into an MLP with one hidden layer:

$$e_a = W_a x_a + b_a \quad e_1 = W_f x_1 + b_f \quad e_2 = W_f x_2 + b_f \tag{4.1}$$

$$h = \text{ReLU}(W_h[e_a e_1 e_2] + b_h) \tag{4.2}$$

$$\hat{y} = W_y h + b_y \tag{4.3}$$

where x_a, x_1, x_2 and e_a, e_1, e_2 are the one-hot vectors and embeddings representing the axis cue, face 1, and face 2, respectively, h is the hidden representation of the MLP, and $\hat{y}$ is the output. Brackets denote concatenation, and ReLU() is the rectified linear unit activation function.

4.3.0.2 Prefrontal Cortex for Cognitive Control

In further simulations the base MLP was augmented with a PFC layer that received the context as input and controlled the units in the hidden layer of the MLP with a gating mechanism:

$$g = c \odot h \tag{4.4}$$

where c is a control signal vector generated from the axis cue, and $\odot$ signifies element-wise multiplication. The output layer of the MLP then acted on the gated hidden layer, rather

than the hidden layer itself (replacing equation 4.3 above):

$$\hat{y} = W_y g + b_y \tag{4.5}$$

Note the PFC was responsible for modulating activity according to the current context, as the MLP no longer received e_a as input. This mechanism is largely consistent with classic neural network models of cognitive control (Cohen et al., 1990; Miller & Cohen, 2001; Rougier et al., 2005), which emphasize the role of the PFC in modulating and regulating the flow of activity in posterior areas through top-down attentional control according to the current goal.

The control signal was determined from the axis cue according to a simple scheme: half of the units in the hidden layer were gated in response to one of the cues, and the other half of the units were gated in response to the other cue.

$$c = \begin{cases} [11...100...0] \cdot \gamma & \text{if axis} = 0 \\ [00...011...1] \cdot \gamma & \text{if axis} = 1 \end{cases} \tag{4.6}$$

where γ determines the strength of the control signal's influence on the hidden units. Note that there was no learning in the PFC: in this work we were interested in the effects of cognitive control and gating on learning in the MLP when trials were blocked or interleaved. Future work will explore methods for introducing learning into the PFC (Flesch et al., 2022b; Tsuda et al., 2020; Wang et al., 2018).

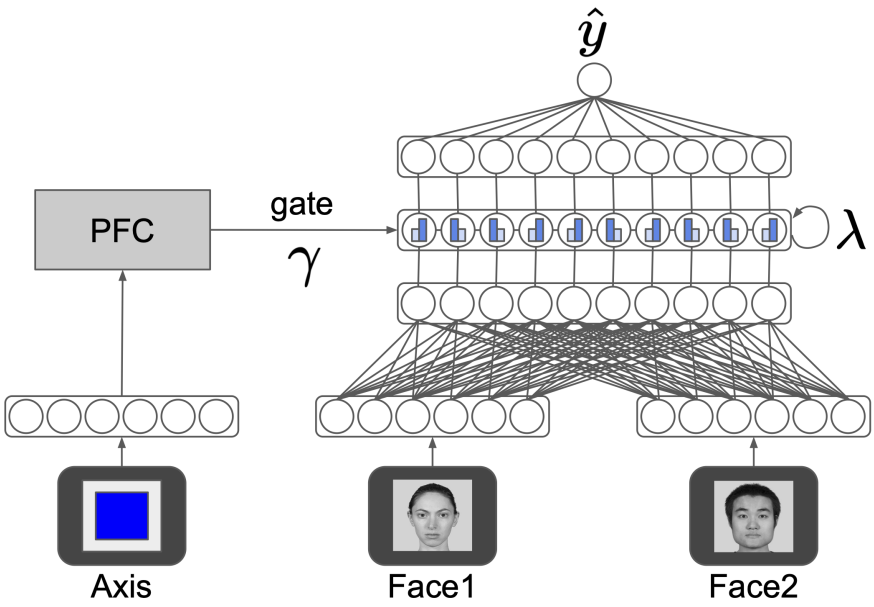


Figure 4.2: Model architecture. The model was trained to predict which of two faces ranked higher on the cued dimension ("Axis"). Inputs were embedded and passed through an MLP. The units in the hidden layer were modulated by a PFC module, which could gate them via element-wise multiplication by numbers from 0 to 1 (shown for illustration purposes as unit-wise probabilities of gating vs. not gating). Additional parameters γ and λ determined the strength of the control signal and the active maintenance, respectively.

4.3.0.3 Active Maintenance

We also implemented a parameter λ that controlled a default bias in the PFC layer to maintain its activity over time:

$$s^{(t)} = \sigma(c^{(t)} + \sigma(s^{(t-1)} - 1 + \lambda)) \tag{4.7}$$

where t indicates time, λ determines the degree to which the previous control signal is added to the current one on each time step, σ is a rectified linear function that returns 0 for inputs less than 0 and 1 for inputs greater than 1, and now the new variable s integrates the control signal over time and acts on the hidden state of the MLP (replacing Equation 4.4):

$$g^{(t)} = s^{(t)} \odot h^{(t)} \tag{4.8}$$

The maintenance parameter (λ) allowed us to control the degree to which the control signal was biased to maintain its activity over time, which introduces a cost when the context (i.e., the axis cue) was switched from trial to trial due to interference from the previous control signal ($s^{(t-1)}$). The bias to actively maintain patterns of activity in PFC is well established (O'Reilly & Frank, 2006), and is fundamental to the important role the PFC plays in working memory, executive functioning, and planning. We hypothesized that these dynamics would be relevant to our setting because when trials are interleaved the switch cost may have negative effects on learning. We used a particularly simple implementation to capture this basic dynamic, but future work will investigate whether its effects on learning play out in more realistic implementations (O'Reilly & Frank, 2006).

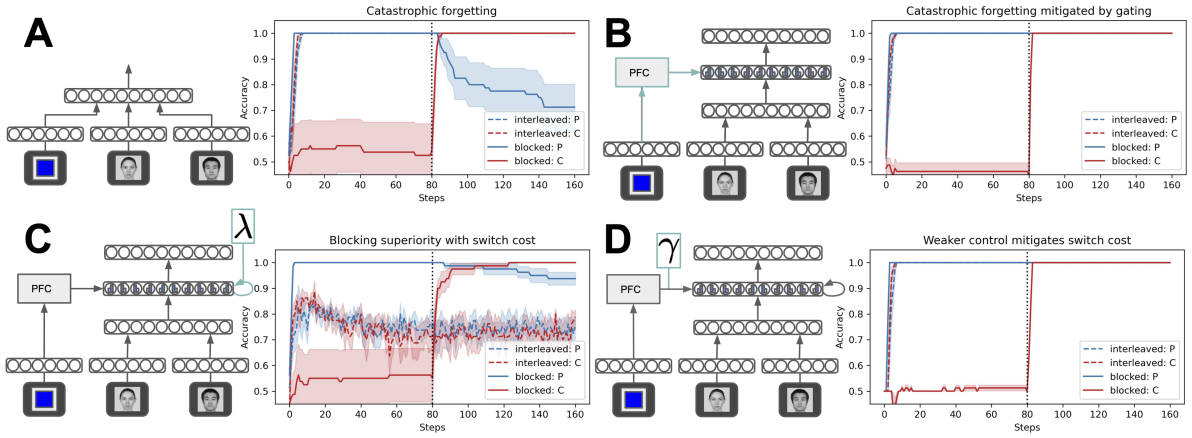


Figure 4.3: Accuracy results. Each plot shows accuracy (y-axis) over the course of training steps (x-axis) for a configuration of the model, depicted by a diagram next to the associated plot. In each experiment, trials were split on the test set by the relevant axis cue (popularity = P, shown in blue, and competence = C, shown in red), and accuracy was measured separately for each in order to show the effects of blocking. Each simulation included 5 runs where trials were interleaved (dashed lines) and 5 runs where trials were blocked (solid lines). Solid areas show SEM across runs. **A)** Catastrophic forgetting occurred in the base MLP model when it was trained on the blocked condition. **B)** Catastrophic forgetting was alleviated by the addition of a control signal from the PFC module (highlighted in cyan). **C)** The model's performance on the interleaved condition suffered when a default active maintenance was introduced in the control signal (shown by self-connection highlighted in cyan), inducing a cost to switching between contexts. **D)** This switch cost was eliminated when the control strength was reduced (shown by γ highlighted in cyan), demonstrating the tradeoff between control strength and switch cost.

4.3.0.4 Implementation Details

Models were built in PyTorch, and were supervised on correct responses with a cross entropy loss function. Models were optimized using backpropagation and Adam (Kingma & Ba, 2015) with a learning rate of 0.001. Embedding vectors had 32 dimensions, and there were 128 units in the hidden layer. For each simulation, 5 runs with different random initializations were performed.

4.4 Results

All versions of the model were trained on both blocked and interleaved conditions. In particular, we explored our hypotheses by testing the model with different configurations of the parameters described above. Accuracy on the test set was evaluated for each social dimension separately in order to assess forgetting in the blocked condition.

4.4.1 Catastrophic Forgetting when Trials are Blocked

First, we reproduced catastrophic forgetting in the model by training the base MLP (without a PFC) on both the blocked and interleaved conditions of the task (see Figure 4.3A). When trials were interleaved, the base MLP model had no problem learning the task, and quickly achieved 100% accuracy on the test set. However, when trials were blocked, we observed catastrophic forgetting: after initially performing well on the first block, over the course of the second block performance progressively declined, indicating increasing forgetting of the relationships along the first dimension that were learned in the preceding block. This result can be understood in the context of CLS theory (McClelland et al., 1995), which suggests that catastrophic forgetting occurs whenever overlapping patterns interfere with each other.

4.4.2 Cognitive Control Mitigates Forgetting

To establish that gating in the PFC can mitigate interference and reduce catastrophic forgetting, we trained the model equipped with a PFC on the same set of conditions (see Figure 4.3B). For the purposes of this experiment, we removed the internal dynamics of the PFC, setting the λ parameter to 0 (no maintenance) and the γ parameter to 1.0. When this model was trained on the task, its performance on interleaved trials was unaffected, and quickly rose to 100% accuracy. However, when it was trained on blocked trials, the catastrophic forgetting observed in the previous experiment was alleviated, and the model was capable of retaining what it had learned in the first block through the subsequent block.

This finding is consistent with the basic principles of CLS (McClelland et al., 1995): when the overlap between patterns of activity in the hidden layer is reduced, interference and forgetting are alleviated. However, CLS theory holds that the hippocampus reduces overlap in its representations with mechanisms that promote sparsity, whereas here we show that a PFC equipped with a dynamic gating mechanism can accomplish a similar goal. This is consistent with the results of previous computational models (Rougier et al., 2005; Tsuda et al., 2020) showing that adaptive gating can offer an alternative mechanism for reducing the overlap between patterns of activity, thereby reducing interference and forgetting.

4.4.3 Blocking Advantage with a Switch Cost

The results above and the results of previous models (Rougier et al., 2005; Tsuda et al., 2020) show that catastrophic forgetting can be reduced when learning occurs in non-overlapping patterns of activity across a layer, thereby explaining the reduced effects of interference observed in humans and other animals as compared with standard neural network models. However, in certain cases human performance has been shown to be *superior* when trials are blocked compared with when they are interleaved (Carvalho & Goldstone, 2014; Flesch et al., 2018; Noh et al., 2016). We hypothesized that this reversal of the catastrophic forgetting phenomenon may be due to the internal dynamics of cognitive control processes (Flesch et al., 2022b), and in particular due to the bias in neurons in the PFC to actively maintain their activity over time (O'Reilly & Frank, 2006). To explore this hypothesis, we implemented a control model with simple recurrent dynamics (see Equation 4.7), keeping the γ parameter at 1.0 but setting the λ parameter to 0.9 (i.e., 90% of the previous control signal is maintained at each time step). The resultant dynamics can be thought of as exhibiting a switch cost (Blackwell et al., 2014; Hyafil et al., 2009), wherein rapidly switching the context or goal (in this case the relevant social dimension) introduces interference due to the ongoing maintenance of the previous context. Note that the cognitive cost of task switching is usually measured in increased reaction times or errors, but here we study it in the context of its

effects on learning.

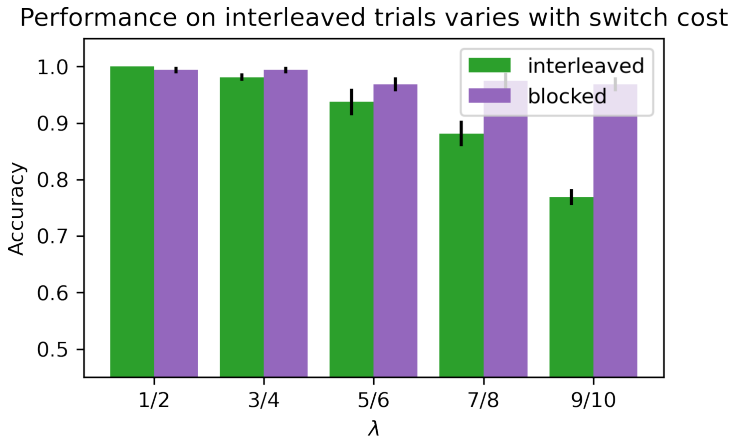


Figure 4.4: Effect of maintenance parameter (λ) on performance. In the blocked condition (purple), accuracy on the test set does not depend much on maintenance. However, as maintenance increases, the cost to switching worsens and performance in the interleaved condition declines.

When these dynamics were introduced, the model was relatively unaffected when trials were blocked, but exhibited a consistent reduction in performance when trials were interleaved (see Figure 4.3C). When trials were interleaved, many switches between contexts occurred throughout training, thereby introducing interference in the control signal, causing processing to be ineffectively modulated according to the current context. We also performed simulations where we systematically varied the λ parameter (see Figure 4.4), showing consistent reductions in performance on the interleaved condition with increased active maintenance.

4.4.4 Tradeoff between Control Strength and Switch Cost

Previous work has suggested a natural tradeoff between the strength of cognitive control and the cost incurred when a context or task-set is switched (Herd et al., 2014): stronger control would be more effective in coordinating activity in other brain regions according to the current goal, but may make rapid switching between task sets or goals more difficult.

To demonstrate this tradeoff, we tested a model with the maintenance (λ) kept at 0.9, but reduced the value of γ (control strength) to 0.1. In this case, the model still performed well when trials were blocked, but the reductions in performance when trials were interleaved disappeared (see Figure 4.3D). This shows that weakening the control signal can reduce the switch cost, aiding performance when there are many switches. Our results are consistent with a tradeoff between the strength of control and the switch cost: without control, catastrophic forgetting is detrimental to performance when trials are blocked, but when control is too strong, interference hurts performance when trials are interleaved.

4.4.5 Analysis of Learned Representations

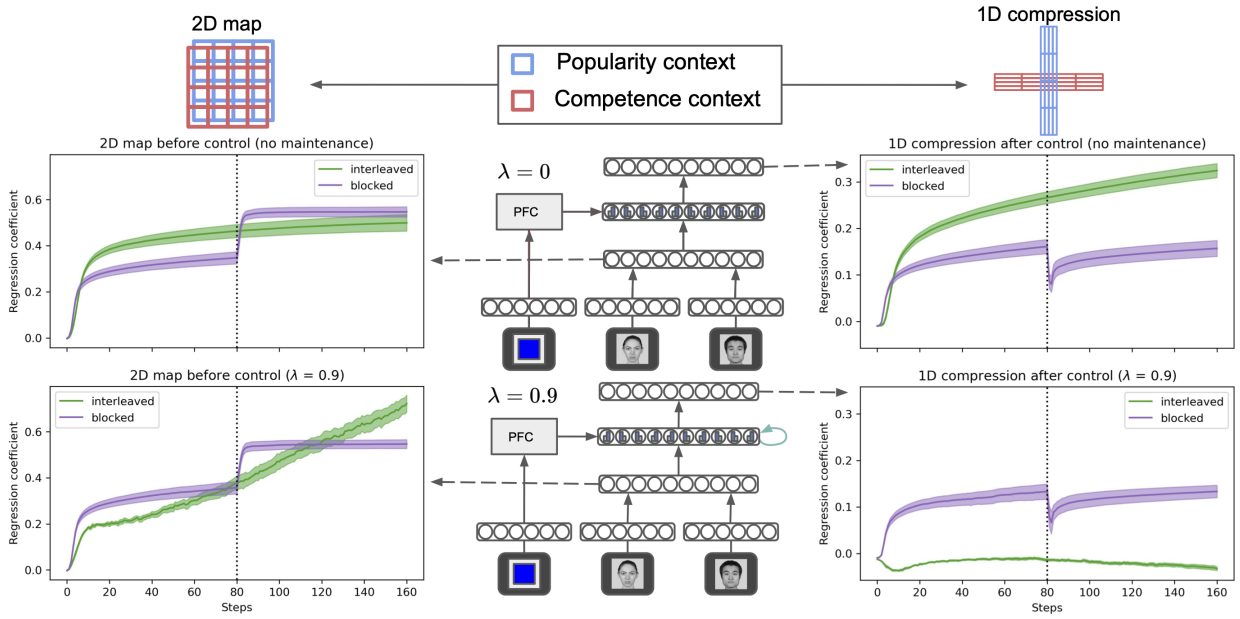


Figure 4.5: Results of analyzing the learned representations of the model. Representations were analyzed in terms of how well they captured the 2D structure of the 4x4 grid (left) and how much the irrelevant dimension of the grid was compressed on each trial (right). Idealized grids depicting these two predictions are shown on the top, where the red grid indicates idealized spacing between representations extracted during trials on which competence was cued, and the blue grid indicates the same for popularity trials. Plots show the beta coefficients over training from performing the relevant regressions. These were conducted on hidden representations either before (left) or after (right) control was applied, on two configurations of the model - one with $\lambda = 0$ (no maintenance) and one with $\lambda = 0.9$. Regression results revealed strong 2D map-like structure in the hidden layer before control was applied, and strong 1D compression of the irrelevant dimension after control was applied. However, when the active maintenance was too strong ($\lambda = 0.9$), the compression effect disappeared in the interleaved condition, indicating a failure to modulate representations according to the current context. Vertical lines indicate the switch in the blocked condition.

The grid structure of our task allowed us to make concrete predictions about the representations that would be learned in the hidden layers of the network (Park et al., 2020; Russin et al., 2021). In particular, we tested whether the model formed 2D map-like representations that captured the basic structure of the grid (Constantinescu et al., 2021; Park et al.,

2020, 2021), and whether these 2D map-like representations were modulated by the current context. Previous work has shown that on a similar task, 2D structure was modulated by the current context, compressing the irrelevant dimension (Flesch et al., 2022a).

Figure 4.5 shows the results of performing a regression on the representations from the hidden layer with hypothetical distance matrices (depicted as idealized map-like representations) as the predictors. We compared the results of this regression throughout training when the maintenance parameter (λ) was set to 0 and 0.9, and when the hidden representations were extracted before and after the control signal was applied (i.e., g and h in equation 4.8). γ was fixed at 1.0.

The model reliably learned the 2D structure of the grid in its hidden representations regardless of the maintenance, as can be seen in the results from the hidden representations before the control signal was applied. This 2D structure was modulated by the current control signal, which had the effect of compressing the currently irrelevant dimension (or equivalently, expanding the relevant dimension). This suggests that the effect of the control signal was to allow the model to generate its response based on the relevant dimension, and to appropriately facilitate learning in the neurons coding for that dimension. However, when trials were interleaved and maintenance (λ) was set to 0.9, the model did not show this compression pattern after control was applied, indicating a failure to modulate its representations according to the current context. This confirmed the idea that the poor performance on interleaved trials when the switch cost was high (see Figure 4.3C) was caused by interference in the control signal.

4.5 Discussion

The neural networks driving current ML research do not perform well in continual-learning settings where incoming data is blocked or otherwise correlated over time (Hadsell et al., 2020). Humans do not exhibit the catastrophic forgetting that plagues these neural networks

in these settings (McClelland et al., 1995), and in some cases even show a learning advantage when trials are blocked (Carvalho & Goldstone, 2014; Flesch et al., 2018). In this work, we built on previous computational frameworks (Flesch et al., 2018; Rougier et al., 2005; Tsuda et al., 2020), and investigated the potential for cognitive control mechanisms in the PFC to induce non-overlapping patterns of activity in order to mitigate interference. Consistent with previous studies (Tsuda et al., 2020), our simulations suggest that these mechanisms can aid learning when trials are blocked over time.

In addition to pattern-separation mechanisms in the hippocampus proposed in CLS (McClelland et al., 1995), and the gating mechanism in PFC proposed here and elsewhere (Rougier et al., 2005; Tsuda et al., 2020) a number of alternative mechanisms for alleviating catastrophic forgetting in neural networks have been explored (Flesch et al., 2018; Kirkpatrick et al., 2017; Velez & Clune, 2017). In particular, Flesch et al. (2018) show that forgetting was reduced on a similar task when their network was augmented with a good inductive prior. However, they did not show an advantage to blocking over interleaving, although they observed this effect in their human experiments. While our approach is not incompatible with the idea that good inductive priors can mitigate catastrophic forgetting, we also show that a bias to maintain activity in the PFC leads to an advantage of blocking over interleaving, providing an explanation for some of the results observed by Flesch et al. (2018) and others.

In work developed concurrently with ours, Flesch et al. (2022b) show an advantage of blocking in a model based on very similar principles. In their framework, a neural network equipped with a context-gating mechanism was modified to have "sluggish" units that maintain information from previous trials, inducing a switch cost that degrades performance when trials are interleaved. Although there were some slight differences in implementation and in interpretation, we believe the broad convergence between this work and ours highlights the potential of these principles for explaining the advantage of blocking observed in humans.

The advantage of blocking can seem to contradict the well-established principles of CLS

(McClelland et al., 1995). However, we show here that a "cortex-like" neural system equipped with mechanisms for cognitive control and active maintenance can enter a different regime than those typically considered in the CLS framework, wherein a reliance on control exposes the system to interference in the control signal caused by rapid context switches. We speculate that in the brain, pattern-separation mechanisms in the hippocampus are usually sufficient to ensure effective learning regardless of whether experiences are correlated over time, but animals such as humans that rely heavily on cognitive control may in some cases *require* learning experiences to be correlated over time due to the bias for active maintenance in the cognitive controller. In our simulations, we introduced this bias to show how it could lead to a learning advantage of blocking, but of course there was no real need for active maintenance in the task (as shown by the good performance of the base MLP when trials were interleaved). We expect that there are good computational reasons that the PFC would have a bias to maintain its activity over time (e.g., related to its role in working memory and planning), and that these may be unrelated to the demands of this particular task. We leave it to future work to show that a system augmented with cognitive control and active maintenance is superior in an absolute sense to one without these mechanisms.

Our simulations were also inspired by the idea that active maintenance engenders a cost to switching between contexts, which must be traded off against the strength with which control can be applied (Herd et al., 2014). The presence of this tradeoff means the cognitive system as a whole must optimize the strength of its control signal according to constraints imposed by learning as well as the current need for control (Shenhav et al., 2013). This optimization may have taken place over the course of evolution (Herd et al., 2014), but it may also occur in real time according to the task at hand (O'Reilly et al., 2020).

Representational analyses showed that cognitive control can act on 2D map-like representations to modulate them according to the current context by compressing irrelevant dimensions and allowing learning to take place in non-overlapping patterns. However, a strong bias to maintain activity over time leads to interference in the control signal, reduc-

ing this effect and leading to poor performance when trials are interleaved. Flesch et al. (2022a) also showed compression along currently irrelevant dimensions in representations of a neural network trained on a similar task. In particular, this occurred in a "rich" regime when their neural network was initialized with small weights. The default random initializations we used in our model were likely small enough to put them in the "rich" regime, but future work will assess the extent to which our results depend on initialization.

Intelligent systems should be capable of continually learning in settings where data is not independently sampled over time. Our simulations demonstrate computational principles that may underlie human continual learning, and help to explain behavioral phenomena observed in human experiments.

4.6 Acknowledgments

We would like to thank the members of the Computational Cognitive Neuroscience and Learning and Decision Making labs at UC Davis, as well as reviewers for helpful comments and discussions. The work was supported by: ONR grants ONR N00014-20-1-2578, N00014-19-1-2684 / N00014-18-1-2116, N00014-18-C-2067, as well as NSF CAREER Award 1846578, and NIH R56 MH119116.

Chapter 5

Complementary Structure-Learning Neural Networks for Relational Reasoning[1]

Jacob Russin*[2], Maryam Zolfaghar*[2], Seongmin A. Park[3], Erie Boorman[3], Randall C. O'Reilly[2] (* denotes equal contribution)

5.1 Abstract

The neural mechanisms supporting flexible relational inferences, especially in novel situations, are a major focus of current research. In the complementary learning systems framework, pattern separation in the hippocampus allows rapid learning in novel environments, while slower learning in neocortex accumulates small weight changes to extract systematic structure from well-learned environments. In this work, we adapt this framework to a task from a recent fMRI experiment where novel transitive inferences must be made according to implicit relational structure. We show that computational models capturing the basic cognitive properties of these two systems can explain relational transitive inferences in both familiar and novel environments, and reproduce key phenomena observed in the fMRI experiment.

[1*] This chapter was originally accepted for publication in the Proceedings of the 43rd Annual Meeting of the Cognitive Science Society (CogSci 2021) (Russin et al., 2021). The opinions expressed here are solely those of the author and do not necessarily reflect the official views of the conference, workshop, or publisher. The original version can be accessed online at: `https://arxiv.org/abs/2105.08944`.

[2] Center for Neuroscience, University of California, Davis

[3] Center for Mind and Brain, University of California, Davis

5.2 Introduction

Humans and non-human animals are capable of navigating efficiently in both novel and familiar environments. For example, in a well-learned environment like one's hometown, it is easy to navigate to new goal locations and plan novel routes. When traveling in a new city, it is also possible to navigate to a novel location by reasoning over recent experiences — even those accumulated on the same day. In both cases, efficiency requires processes or representations that allow generalization beyond previous experience. This kind of generalization has been a long-standing issue in cognitive science, and was integral to early arguments against behaviorism, where it was claimed that a simple stimulus-response mapping could not account for such behaviors (Tolman, 1948).

More recent work has investigated the computational and neural mechanisms underlying cognitive maps, or representations that capture the structure of the environment and thereby support generalization (Park et al., 2020; Whittington et al., 2020; Behrens et al., 2018). This work has emphasized the importance of certain neocortical areas such as the entorhinal cortex (EC) for spatial reasoning and vector-based navigation (Moser et al., 2008). Furthermore, it has been argued that these structured spatial representations may be leveraged for other kinds of abstract relational reasoning in humans (Behrens et al., 2018). Relatedly, although neural networks have enjoyed massive success on difficult machine-learning tasks in recent years these models are known to fail on out-of-distribution or extrapolation problems (Lake et al., 2017) such as those requiring transitive inferences.

Here, we apply the well-supported complementary learning systems (CLS) framework (McClelland et al., 1995; O'Reilly et al., 2011) to explore two qualitatively different neural mechanisms underlying spatially-grounded relational reasoning abilities in novel and familiar environments. The CLS framework has emphasized the computational justification for learning mechanisms unfolding on two different timescales, as supported by separate brain areas. Slow learning in neocortex allows for the development of more abstract representations that integrate across many experiences and can be leveraged to make novel inferences.

However, this kind of learning is not possible in naturalistic environments where sequences of events are not presented in an interleaved or random order, as when one explores only one part of an environment at a time. This is due to the well-known *catastrophic forgetting* phenomenon, where previous learning is erased by new experiences when learning occurs too quickly or training is not sufficiently interleaved (McClelland et al., 1995). The CLS framework proposes that fast learning can occur in the hippocampus due to its pattern-separated, sparse representations. These representations have little overlap across examples, and therefore allow fast learning of novel episodes, i.e., *episodic memory* (Yonelinas et al., 2019), to occur without catastrophic interference.

In the CLS framework, slow cortical learning is needed to build up structural or relational representations over time, which provide the foundation for systematic inferences. However, for more unfamiliar situations, rapid hippocampal learning is required. Previous work has found evidence suggesting a role for the hippocampus in rapid generalization (Eichenbaum, 2004; Zeithamova et al., 2012), and that a hippocampal model informed by the CLS framework can explain these findings when it is augmented with a recurrent similarity-based computation, proposed to be supported by "big-loop" recurrence between the hippocampus and the neocortex and within the hippocampus itself (Kumaran & McClelland, 2012).

Here, we build on this work and investigate the interplay between slow generalization in neocortex and rapid generalization in the episodic memory system with computational models based on the principles of the CLS framework. Our model of the episodic memory system is similar to previous work (Kumaran & McClelland, 2012) in that it allows rapid generalization in unfamiliar environments, but relies on different computational mechanisms to do so (see Discussion). Our models of the cortical system and the episodic memory system were both tested on a novel non-spatial structure-learning paradigm from a recent fMRI experiment (Park et al., 2020). Importantly, the task required transitive inferences based on learning over two different timescales: training experience over multiple days, and training examples given on the same day as the inference test. In the following, we briefly

outline the key findings of the experiment and offer a conceptual framework that integrates them with the CLS perspective. We then describe the computational models that were built to capture the basic properties of the proposed conceptual framework, and show that these models are capable of performing transitive inferences in the same task and reproduce other key findings from Park et al. (2020).

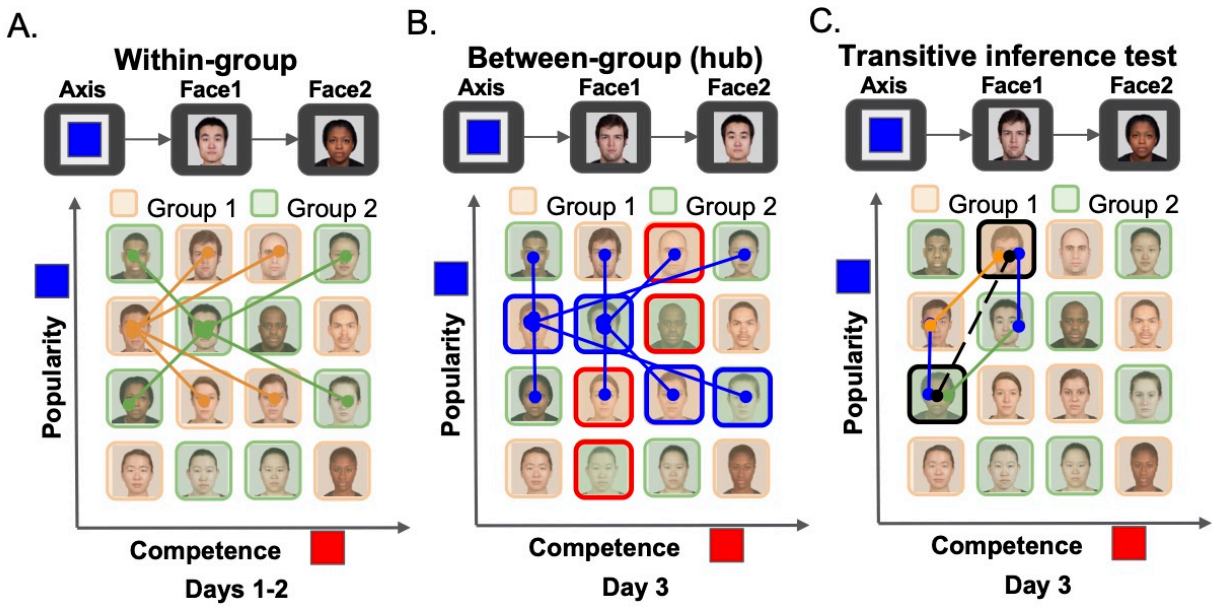


Figure 5.1: Experimental paradigm used in Park et al. (2020). Participants learned the relative ranks of pairs of faces on an implicit grid with two axes: competence and popularity. These faces were split into two groups (shown in green and orange). **A)** Over the first two days, participants were trained on within-group pairs that differed by a rank of 1 along each designated axis. Example pairs on the popularity axis are shown with orange lines (within group 1) and green lines (within group 2). **B)** On the third day, participants learned between-group pairs containing exactly one hub linking the two groups. There were a total of 8 hubs, and each was associated with a certain axis (shown by red and blue outlines). Each hub was paired with the 4 faces from the other group that differed by a rank of 1 along the designated axis. Examples of such pairs are shown for two hubs with blue lines (indicating the popularity axis). **C)** The third day included the fMRI experiment and transitive inference test. Participants were tested on pairs of faces from different groups that could be connected through one of the two hubs on the appropriate axis. Green, orange, and blue lines indicate the training pairs (2 within-group and 2 between-group, which are shown in A and B) that could be used to make the transitive inference for the pair indicated by the black dotted line.

5.2.1 fMRI Experiment

Park et al. (2020) studied the neural mechanisms underlying transitive inference performance on the structure-learning task illustrated in Figure 5.1. Participants learned to make judgments about the "popularity" or "competence" of 16 people through pair-wise comparisons along one of these two axes at a time. Unknown to the participants, these 16 faces were arranged in a 4x4 2D grid, and were implicitly separated into two groups. In the first two days of training participants only learned about within-group pairs that different by a rank of 1 (see Figure 5.1A). On the third day of the experiment, participants learned about between-group pairs containing certain faces that acted as hubs between the two groups (see Figure 5.1B). This training provided sufficient evidence to allow participants to integrate their previously separated cognitive maps, but was conducted on the same day as fMRI scanning. In the scanner, participants performed a transitive inference test in which unseen pairs of faces from different groups were compared (see Figure 5.1C). For each of these test pairs, one of two corresponding hubs could be used to make the transitive inference. The results we focused on in our work can be summarized as follows:

1. Participants exhibited good transitive inference performance, achieving 93.6% mean accuracy on the unseen pairs tested in day 3.
2. Map-like representations were found in several brain areas, including ventromedial prefrontal cortex (vmPFC) and entorhinal cortex (EC). Patterns of activity in these areas demonstrated sensitivity to the ground-truth Euclidean distances between faces in the implicit grid. However, these effects were significantly reduced when the analysis was restricted to between-group pairs that were not encountered during training.
3. A repetition-suppression analysis in hippocampus suggested that one of the two relevant hubs was retrieved from episodic memory at the time of inference.

Taken together, these findings suggest that cortical learning systems in vmPFC and EC were able to integrate across the pairs of faces encountered during training to form map-

like representations that would be useful for making transitive inferences within groups. However, the effects in these areas were reduced when the analysis was restricted to novel between-group pairs, and participants seemed to retrieve the relevant hubs from episodic memory in hippocampus during the transitive inference test. Thus, although the within-group pairs were well-learned over the first two days of training, these groups may not have been fully integrated into a single coherent cognitive map at the time of testing. This may have forced participants to rely instead on hippocampal retrieval of recently-learned between-group training episodes (which always included a hub) to generalize during the transitive inference test. Thus, there appear to be two separable cognitive mechanisms that allow for relational transitive inferences to be made in this task: 1) if given enough training time, cortical areas such as vmPFC and EC can learn representations that reflect the implicit relational structure of the grid, and 2) an episodic retrieval mechanism can ensure good transitive inference performance with pairs that were seen only on the same day as the test. Below we outline a general framework that integrates these findings, and the apparent redundancy in these two systems, with the CLS perspective.

5.3 Complementary Structure-Learning Systems

The CLS framework explains how the brain can support integrative representation learning without suffering from catastrophic forgetting (McClelland et al., 1995; O'Reilly et al., 2011). However, the CLS framework also emphasizes other important reasons for fast learning in an episodic memory system. In particular, slow cortical learning may be insufficient to allow for efficient adaptation in relatively unfamiliar environments (Kumaran & McClelland, 2012). The findings from Park et al. (2020) suggest that humans are capable of making novel transitive inferences using experiences acquired on the same day. Furthermore, they show that these inferences are mediated by hippocampal retrieval of the intermediate states (i.e., hubs) that would allow such inferences to occur. Taken together, these findings suggest

that the dual-process view emphasized in CLS may explain the apparent redundancy in structure-learning mechanisms studied in neuroscience and psychology (see Table 5.1).

Table 5.1: Complementary structure-learning systems.

System	Properties
Cortical learning	• Learns slowly through small, incremental weight changes • Inference is fast and less effortful with map-like representations
Episodic memory	• Learning can be fast due to sparse, pattern-separated representations • Inference is slower, requiring cognitive control for deliberate, goal-directed retrieval

In the case of spatial navigation, slow cortical learning can integrate across many experiences to form map-like representations. This system is capable of directly utilizing its integrative representations without further processing, and can thus make inferences rapidly. However, this system would not be able to make inferences in a newly learned environment if it did not have time to integrate across particular episodes (Kumaran & McClelland, 2012). This may have been the case in the transitive inference test conducted on the same day as the between-group training in the fMRI study (Park et al., 2020). Fast episodic learning, on the other hand, can immediately store memories of individual experiences, allowing inferences to

be made in unfamiliar environments based on few such experiences. However, the episodic nature of its representations do not allow the sort of direct inferences that are available to the cortical system. Instead, transitive inferences require a slower, more deliberate process of goal-directed retrieval and further processing of the stored memories (Zeithamova et al., 2012). An organism equipped with both systems would be capable of making novel inferences in both familiar and unfamiliar environments. In the following, we provide evidence from models that capture, on a *computational* level, the basic properties of the proposed complementary structure-learning systems, and show that these systems reproduce key findings from Park et al. (2020).

5.4 Modeling Framework

We simulated[4] each of our models on the training and testing procedure used in the task, including its within-group and between-group structure and transitive-inference test. In particular, each trial consisted of a presentation of two faces and the axis along which the judgment should be made (i.e., "competence" or "popularity"). The models were required to make a binary judgment about whether the first face ranked higher or lower than the second face along the specified axis.

[4]All data and code used for experiments and analyses are available at https://github.com/MaryZolfaghar/CSLS

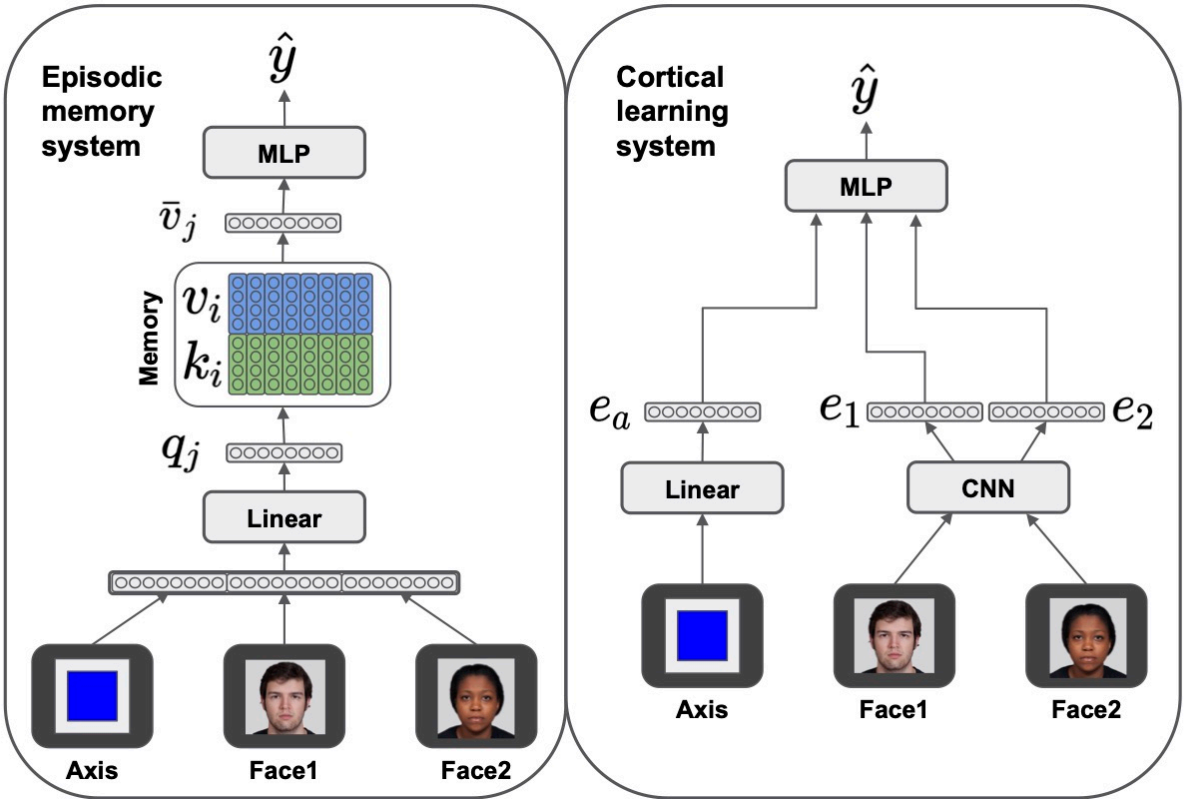


Figure 5.2: Model architecture. (Left) The episodic memory system stores representations of individual training trials in a key-value memory. New inferences are made by querying the memory to retrieve the relevant trials, which are then processed by an MLP to generate an answer. (Right) The cortical learning system was modeled as a simple feedforward network with convolutional layers to process the images. This system relies on its learned representations to perform transitive inferences.

5.4.1 Cortical Map-Building

The cortical representation-learning system should accumulate small updates over many trials to build map-like representations that can be directly utilized to make transitive inferences. We modeled this process with a simple feedforward neural network with two convolutional layers (see right side of Figure 5.2). Face images were taken from the same database used in the fMRI experiment (Strohminger et al., 2016), and were downsampled to 64x64 and grayscaled for faster simulation. The within-group and between-group hub samples were all trained simultaneously (i.e., the pairs that were trained on different days

of the fMRI experiment were trained simultaneously in the model). This is because the purpose of our model of the cortical system was to show that, if given enough training time, it could perform transitive inferences based on its learned representations, and allow fast inference in familiar environments. Each face was processed with the same convolutional neural network, and the axis variable, encoded as a one-hot vector, was embedded with a linear layer. These three embeddings were then concatenated and passed through a multi-layer perceptron (MLP) with rectified linear unit (ReLU) activation functions. This network captures the basic properties of slow cortical learning in that it accumulates small updates to its synaptic weights over many trials, and makes inferences directly based on its learned representations of each face.

5.4.2 Goal-Directed Episodic Memory Retrieval

The episodic memory system should learn quickly by storing individual training episodes, and make inferences by retrieving the previous trials that are relevant to the current one (McClelland et al., 1995; Kumaran & McClelland, 2012). For this purpose, we used a neural memory system (see left side of Figure 5.2) with a soft retrieval mechanism (Botvinick et al., 2019). This memory system immediately stores each trial (x_i) seen during training as a key, value pair: $k_i = W_k x_i$, $v_i = W_v x_i$, where k_i is the key, v_i is the value, and x_i is the trial, which is a concatenation of a one-hot encoding of each face, the axis variable (a), and the correct answer (y) of the ith trial. One-hot encodings were used for faces under the assumption that what is stored in the episodic memory system should be a highly processed, sparse encoding (McClelland et al., 1995). To make an inference, the model generates a query according to the current pair of faces: $q_j = W_q x_j^- + b_q$, where x_j^- indicates the jth test trial with the same components but excludes the correct answer (y). This query is then used to retrieve the memories most relevant to the current trial:

$$\bar{v}_j = \text{softmax}\left(q_j K^T\right) V \tag{5.1}$$

where K and V are matrices containing all of the stored memories. Finally, the retrieved memories $\bar{v}_j$ are passed through an MLP to produce the final answer: $\hat{y}_j = \text{MLP}(\bar{v}_j)$. This network captures the basic properties of a fast-learning episodic memory system in that each training episode can be stored in memory immediately upon presentation, and must later be retrieved in a goal-directed way to make a transitive inference.

An interesting problem in modeling episodic memory concerns the learning mechanisms involved in goal-directed memory retrieval. We assume that the human participants recruited for the Park et al. (2020) study had extensive prior experience with goal-directed memory retrieval and everyday transitive inferences. We therefore adopted a meta-learning strategy (Santoro et al., 2016) to model this prior experience, and pretrained the episodic memory system to learn to solve new transitive inference problems sampled from a distribution of such tasks. This pretraining consisted of slow, incremental changes to the weights responsible for mapping into and out of the episodic memory itself, and should thus be thought of as occurring in memory-related cortical areas rather than in the hippocampus proper (McClelland et al., 1995). The system was pretrained on a distribution that was generated by permuting the positions of each face in the 4x4 grid. For each new task, the memory system stored training samples in its memory and used them to make transitive inferences in the testing phase, where it accumulated errors that were then used to update its learnable parameters. The model was then tested on how well it could generalize with a new configuration of faces it had never seen before.

This kind of meta-learning strategy was adopted from previous work (Lake, 2019), and shares with it the limitation that the pretraining tasks are unrealistically similar to the final test — future work will examine the extent to which the model can generalize when trained on substantially different goal-directed retrieval and transitive inference tasks. Additionally, although the resulting goal-directed retrieval mechanism in this model does not capture the hypothesized properties of being deliberative and requiring cognitive control (thus making inferences slower), a more biologically grounded approach involving frontal cortical executive

function systems, planned for future work, would do so. Our purpose in the current study was to show that this system was capable of making transitive inferences in a structured environment.

5.4.3 Implementation Details

Models were built using PyTorch. Models were trained with a cross-entropy loss function and Adam optimizer (Kingma & Ba, 2015) with a batch size of 32 and a learning rate of 0.001.[1] The cortical system was trained for 100 epochs with a batch size of 32. The axis embedding (e_a) had 32 dimensions. Convolutional layers had no padding, a kernel size of 3, a stride of 2, and 4 and 8 channels in the first and second layers, respectively. Each convolutional layer was followed by a max-pooling layer with a kernel size of 2. The CNN contained a linear layer to produce flat 72-dimensional vectors e_1 and e_2, which were passed to the final MLP, which had 128 hidden units. The episodic memory system was pre-trained on 10,000 permutations. Queries, keys, and values were all 32-dimensional, and the final MLP had 64 hidden units.

5.5 Results

Both systems proved to be capable of performing transitive inferences in the task environment from Park et al. (2020): each system achieved 100% accuracy on the held-out test set in which unseen between-group pairs were tested. This validates the idea that the two qualitatively different kinds of learning system outlined above are capable of reproducing human transitive inference performance on the task. To investigate how these qualitative differences might have affected each model's inference strategy, we performed analogues of key analyses done in the experiment (Park et al., 2020) to interpret the behavior of each system, and to evaluate them against empirical results obtained in the fMRI experiment.

[1]Note that the "learning rate" for the episodic memory refers to the weight updates in the pre-training phase. During training, it immediately stored experiences upon presentation.

5.5.1 Cortical Representations Reflect Task Structure

To understand how the cortical system had learned to represent each of the faces, we conducted analyses on the embeddings of each face obtained from the CNN. Visualization of these embeddings with principal components analysis (PCA; see Figure 5.3) showed that the cortical system had learned to represent the faces in terms of their structured relationships, i.e., it had learned map-like representations. These top two principal components explained 95.1 % of the variance in the embeddings, indicating that the model had learned to represent the faces on a near two-dimensional grid.

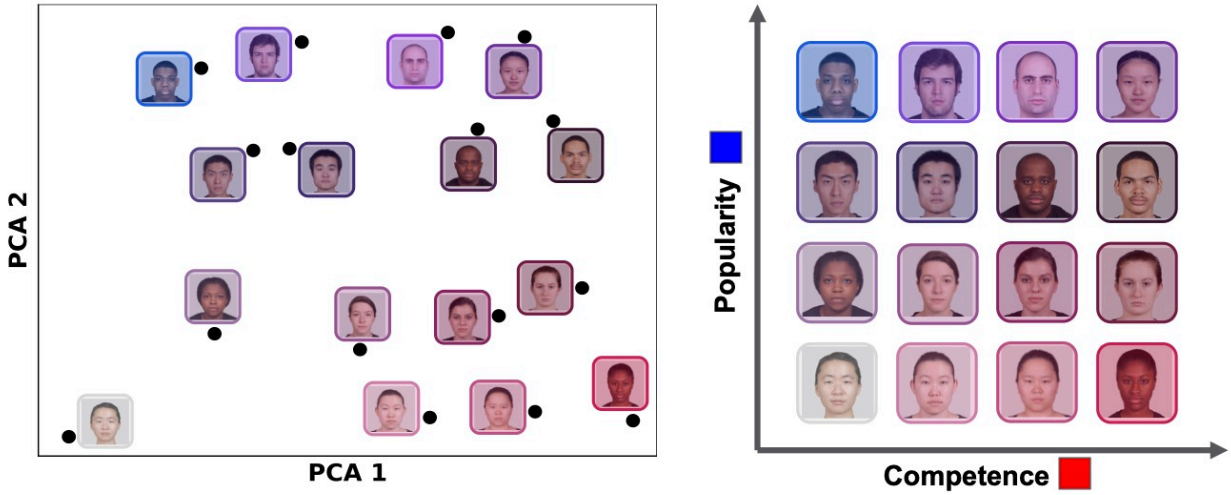


Figure 5.3: Visualization of embeddings learned by the cortical system. Embeddings for each face was projected into two dimensions using PCA, and then rotated by a fixed angle for illustration purposes. The relative positions of the representations indicate that the model has learned to represent the faces in terms of their implicit relational structure.

In addition to the PCA, we conducted an analysis similar to those done in the fMRI experiment (Park et al., 2020), where patterns of activity in vmPFC and EC were found to be sensitive to Euclidean distances in the ground-truth grid. We measured the Pearson correlation between ground-truth Euclidean distances in the grid and the observed distances between each pair of embeddings. A strong correlation was observed ($r(118) = .910$, $p < 0.001$), indicating the same sensitivity to structured relationships in the grid.

5.5.2 Episodic Memory System Retrieves Hubs

In the original fMRI experiment, a repetition-suppression analysis suggested that participants were retrieving the relevant hubs from hippocampus during the transitive inference test (see Figure 5.1C). Although the episodic memory model did not have analogous neural adaptation dynamics that would allow us to model repetition suppression, we conducted an analysis on the retrieved memories to see how the hubs were being used to make transitive inferences. The soft episodic retrieval mechanism shown in equation (5.1) uses a softmax to produce a probability distribution over all of the items in memory. For each test trial, we directly analyzed the weights applied to the memories for the relevant hub trials and compared these weights to the irrelevant memories (see Figure 5.4). Memories were counted as relevant if they included one of the two possible between-group hubs for the given pair of faces, and connected this hub to one of the two faces from the current trial (see Figure 5.1C). This revealed that the weights applied to the relevant hub memories were usually the largest (i.e., the hub trials were retrieved more than the irrelevant trials). Furthermore, an additional analysis found that in every test trial, one of the two possible "paths" connecting the first face to the second face (e.g., in Figure 5.1C, the path through the blue line and green line or the path through the blue line and orange line) was in the top 5% of retrieved memories.

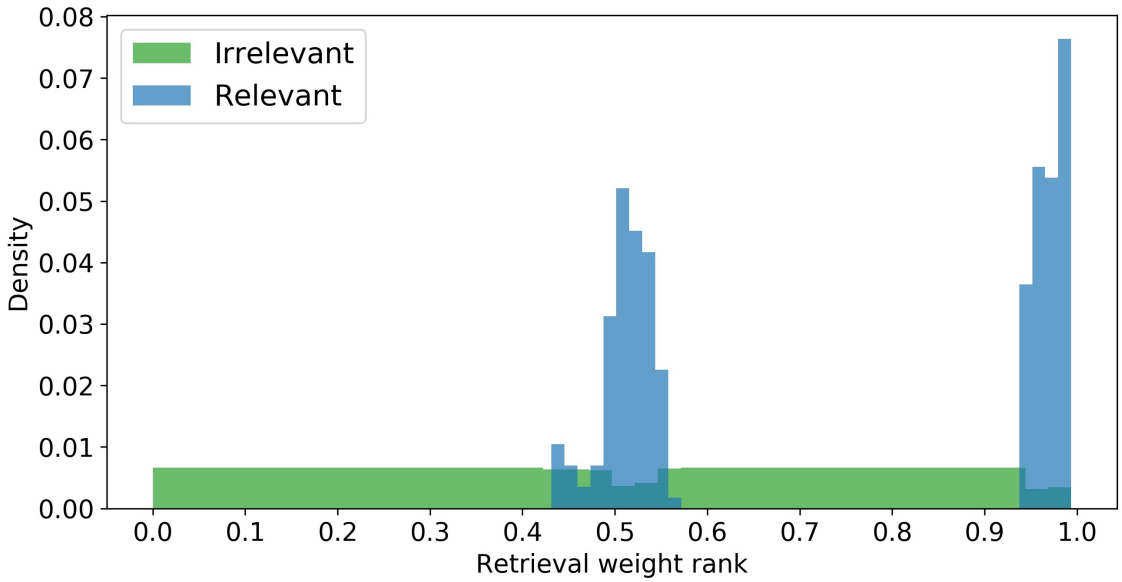


Figure 5.4: Histograms of relevant and irrelevant trials retrieved from the episodic memory during testing. Relevant memories, which always contained a hub, made up the majority of those with the highest weights. This reproduces the fMRI finding that the hubs were retrieved during the inference test. Note that it was not necessary to retrieve every relevant memory to get the correct answer, which may be why the relevant memories were not always retrieved with the highest weights. Counts were normalized to probability densities.

5.6 Discussion

The CLS perspective emphasizes the need for two qualitatively different learning systems in the brain: fast learning can occur in the hippocampus due to its pattern-separated representations, while learning in the neocortex must be slow due to its overlapping representations (McClelland et al., 1995). Here, we investigate this conceptual framework in the domain of structure-learning and relational transitive inference (Kumaran & McClelland, 2012), and propose an analogous distinction. The episodic memory system can learn quickly and generalize in relatively unfamiliar environments, but requires a more deliberate goal-directed retrieval process. The cortical system learns slowly but can make fast inferences in familiar environments from its learned representations. As in the traditional CLS framework, an organism equipped with both systems would retain the benefits of each, allowing generalization

in both novel and familiar environments. Our computational models provide evidence that each of the two proposed systems are able to perform well on a difficult relational transitive-inference test under different circumstances: the cortical system can make these inferences once extensive experience with an environment has been accumulated, while the episodic system can do so quickly, as long as it has had sufficient prior exposure to similar tasks. Our models also reproduce the basic findings from a human fMRI experiment (Park et al., 2020): the cortical system learns map-like representations that encode the implicit relational structure of the grid, while the episodic memory system learns to query its memory for the appropriate hubs connecting the two groups.

Kumaran & McClelland (2012) investigate rapid generalization in a hippocampal model based on the principles of the CLS framework. The model allows retrieval-based inferences to be made — despite the nature of its pattern-separated representations — by incorporating a recurrent similarity computation that can perform associative linking (Eichenbaum, 2004; O'Reilly & Rudy, 2001). This computation is hypothesized to be supported by "big-loop" recurrence (Koster et al., 2018). Our model of the episodic memory system is not inconsistent with hippocampal retrieval-based inferences based on dynamic similarity computation, and in fact the fMRI experiment showed evidence of the presence of such similarity structure in the hippocampus (Park et al., 2020). In addition, the strategy used by our model to solve transitive inference problems appeared consistent with the associative linking exhibited by the model of Kumaran & McClelland (2012), as shown by the retrieval of hubs linking the two groups (see Figure 5.4). However, in our model this strategy emerged over the course of (meta-)learning the structure of transitive inference problems, suggesting a more general mechanism that could be applied to goal-directed retrieval tasks that are not solvable with an associative linking strategy. This learning mechanism has been shown to be useful in the context of one-shot learning (Santoro et al., 2016), and compositional generalization (Lake, 2019). More work is needed to investigate whether hippocampal involvement in rapid generalization occurs when such a strategy is not possible, and whether our model would

benefit from the recurrent computation intrinsic to the model of Kumaran & McClelland (2012).

Our modeling framework shares important properties with the Tolman-Eichenbaum Machine (TEM) (Whittington et al., 2020), which also incorporates meta-learning and models structure-learning in EC. A critical difference between these two models is that in TEM, structure-learning depends on backpropagating error signals through the hippocampus, whereas the CLS framework holds that slow cortical learning can operate independent of the hippocampus to facilitate inferences, consistent with the remarkably intact abilities of early developmental amnesics (Vargha-Khadem et al., 1997).

Our proposed framework integrates ongoing empirical findings about cognitive maps with the CLS perspective, but it also shares some similarities to other prominent dual-process views in cognitive science. For example, prominent theories emphasize a distinction between habitual and controlled processing (O'Reilly et al., 2020), fast and slow thinking (Kahneman, 2011) and model-free and model-based RL (Botvinick et al., 2019). Our conceptual framework proposes a similar distinction between the deliberative, goal-directed retrieval that must occur in the episodic memory system to make transitive inferences, and the more automatic or vector-based generalization that can occur in the cortical system in familiar environments.

There are some important limitations of our current computational models that must be addressed in future work. First, although the two proposed cognitive systems are hypothesized to be realized in the hippocampus and cortical areas such as EC, we have not focused on the interactions that should occur between the two systems. For example, the representations stored in episodic memory should be directly informed by the slowly changing representations learned in cortex, reflecting cortical inputs to the hippocampus. The fMRI study found that map-like representations were also present in the hippocampus (Park et al., 2020), perhaps due to interactions with nearby cortical areas (Kumaran & McClelland, 2012). A more integrated model would show how map-like representations in cortex

can influence hippocampal processing, and how reliance on the episodic memory early in learning shifts to reliance on the cortical system later in learning. This shift may occur due to the cognitive demands imposed on an episodic retrieval mechanism required to reason over individual past experiences. The current episodic memory system does not capture the cognitive control hypothesized to be required for inferences to be made; future work will address this with a more integrated model that deploys an episodic retrieval mechanism with costly sequential processing. Finally, the neural networks used in our models biologically implausible in a number of ways, e.g., the use of a slot-based episodic memory and the standard backpropagation algorithm. Future work will focus on more biologically plausible learning algorithms and more detailed biology of the neocortex and hippocampus.

5.7 Acknowledgments

We would like to thank the members of the Computational Cognitive Neuroscience lab and the Learning and Decision Making lab, as well as reviewers for helpful comments and discussions. The work was supported by: ONR grants ONR N00014-20-1-2578, N00014-19-1-2684 / N00014-18-1-2116, N00014-18-C-2067. J.R. was supported by the NIMH under Award Number T32MH112507. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Aben, B., Calderon, C. B., Van den Bussche, E., & Verguts, T. *2020*, *Cognitive effort modulates connectivity between dorsal anterior cingulate cortex and task-relevant cortical areas*, Journal of Neuroscience, 40, 3838

Alday, P. M. *2019*, *How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting its limits*, Psychophysiology, 56, e13451, doi: 10.1111/psyp.13451

Andersen, P., & Andersson, S. A. 1968, Physiological Basis of the Alpha Rhythm (Appleton-Century-Crofts)

Bae, G.-Y., & Luck, S. J. *2018*, *Dissociable decoding of spatial attention and working memory from EEG oscillations and sustained potentials*, Journal of Neuroscience, 38, 409

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., et al. *2018*, *What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior*, Neuron, 100, 490, doi: 10.1016/j.neuron.2018.10.002

Benjamini, Y., & Hochberg, Y. *1995*, *Controlling the false discovery rate: A practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society. Series B (Methodological), 57, 289, doi: 10.1111/j.2517-6161.1995.tb02031.x

Bernardi, S., Benna, M. K., Rigotti, M., et al. *2020*, *The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex*, Cell, 183, 954, doi: 10.1016/j.cell.2020.09.031

Berry, D. C., & Dienes, Z. 1993, Implicit Learning: Theoretical and Empirical Issues (Hillsdale, NJ: Erlbaum)

Blackwell, K. A., Chatham, C. H., Wiseheart, M., & Munakata, Y. *2014*, *A Developmental Window into Trade-Offs in Executive Function: The Case of Task Switching versus Response Inhibition in 6-Year-Olds*, Neuropsychologia, 62, 356, doi: 10.1016/j.neuropsychologia.2014.04.016

Botvinick, M., Ritter, S., Wang, J. X., et al. *2019*, *Reinforcement Learning, Fast and Slow*, Trends in Cognitive Sciences, 23, 408, doi: 10.1016/j.tics.2019.02.006

Brainard, D. H. *1997*, *The Psychophysics Toolbox*, 10, 433, doi: 10.1163/156856897X00357

Brouwer, G. J., & Heeger, D. J. *2011*, *Cross-orientation suppression in human visual cortex*, Journal of neurophysiology, 106, 2108

Brown, T. B., Mann, B., Ryder, N., et al. *2020*, Language Models Are Few-Shot Learners

Bullier, J. *2001*, *Integrated Model of Visual Processing.*, 36, 96

Calderone, D. J., Lakatos, P., Butler, P. D., & Castellanos, F. X. *2014*, *Entrainment of Neural Oscillations as a Modifiable Substrate of Attention*, Trends in Cognitive Sciences, 18, 300–309

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, 481, doi: 10.3758/s13421-013-0371-0

Cleeremans, A., & McClelland, J. L. *1991*, *Learning the Structure of Event Sequences*, Journal of Experimental Psychology: General, 120, 235, doi: 10.1037/0096-3445.120.3.235

Clegg, B. A., DiGirolamo, G. J., & Keele, S. W. *1998*, *Sequence Learning*, Trends in Cognitive Sciences, 2, 275

Cohen, J. D., Dunbar, K., & McClelland, J. L. *1990*, *On the Control of Automatic Processes: A Parallel Distributed Processing Model of the Stroop Effect*, Psychological Review, 97, 332

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. *2021*, *Organizing Conceptual Knowledge in Humans with a Gridlike Code*, 352, 1464, doi: 10.1126/science.aaf0941

Coomans, D., Deroost, N., Vandenbossche, J., Van den Bussche, E., & Soetens, E. *2012*, *Visuospatial Perceptual Sequence Learning and Eye Movements*, 59, 279, doi: 10.1027/1618-3169/a000155

Daltrozzo, J., & Conway, C. M. *2014*, *Neurocognitive Mechanisms of Statistical-Sequential Learning: What Do Event-Related Potentials Tell Us?*, Frontiers in Human Neuroscience, 8, doi: 10.3389/fnhum.2014.00437

de Cheveigné, A., & Nelken, I. *2019*, *Filters: when, why, and how (not) to use them*, Neuron, 102, 280

Delorme, A., & Makeig, S. *2004*, *EEGLAB: an open-source toolbox for analysis of single-trial EEG dynamics including independent component analysis*, Journal of neuroscience methods, 134, 9

Dennis, N. A., Howard, J. H., & Howard, D. V. *2006*, *Implicit sequence learning without motor sequencing in young and old adults*, Experimental brain research, 175, 153

Deroost, N., & Soetens, E. *2006*, *Spatial processing and perceptual sequence learning in SRT tasks*, Experimental psychology, 53, 16

Destrebecqz, A., & Cleeremans, A. *2001*, *Can Sequence Learning Be Implicit? New Evidence with the Process Dissociation Procedure*, 8, 343, doi: 10.3758/BF03196171

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. *2019*, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in Proc. of the 2019 Conf. of the NA Chapt. of the Assoc. for Comp. Ling., ed. J. Burstein, C. Doran, & T. Solorio (Minneapolis, MN, USA: Association for Computational Linguistics), 4171–4186

Drisdelle, B. L., Aubin, S., & Jolicoeur, P. *2017*, *Dealing with ocular artifacts on lateralized ERPs in studies of visual-spatial attention and memory: ICA correction versus epoch rejection*, Psychophysiology, 54, 83

Eichenbaum, H. *2004*, *Hippocampus: Cognitive Processes and Neural Representations That Underlie Declarative Memory*, Neuron, 44, 109, doi: 10.1016/j.neuron.2004.08.028

Ester, E. F., Anderson, D. E., Serences, J. T., & Awh, E. *2013*, *A neural measure of precision in visual working memory*, Journal of cognitive neuroscience, 25, 754

Fahrenfort, J. J., Grubert, A., Olivers, C. N., & Eimer, M. *2017*, *Multivariate EEG analyses support high-resolution tracking of feature-based attentional selection*, Scientific reports, 7, 1886

Fiebelkorn, I. C., & Kastner, S. *2019*, *A Rhythmic Theory of Attention*, 23, 87, doi: 10.1016/j.tics.2018.11.009

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. *2018*, *Comparing Continual Task Learning in Minds and Machines*, Proceedings of the National Academy of Sciences, 115, E10313, doi: 10.1073/pnas.1800755115

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. *2022a*, *Orthogonal Representations for Robust Context-Dependent Task Performance in Brains and Neural Networks*, Neuron, S0896, doi: 10.1016/j.neuron.2022.01.005

Flesch, T., Nagy, D. G., Saxe, A., & Summerfield, C. *2022b*, *Modelling Continual Learning in Humans with Hebbian Context Gating and Exponentially Decaying Task Signals*, arXiv:2203.11560 [cs, q-bio]. http://ascl.net/2203.11560

Foerde, K., & Poldrack, R. A., Procedural Learning in Humans. 2016, in Reference Module in Biomedical Sciences (Elsevier)

Foster, J. J., Sutterer, D. W., Serences, J. T., Vogel, E. K., & Awh, E. *2017*, *Alpha-band oscillations enable spatially and temporally resolved tracking of covert spatial attention*, Psychological science, 28, 929

Frensch, P. A., Lin, J., & Buchner, A. *1998*, *Learning versus Behavioral Expression of the Learned: The Effects of a Secondary Tone-Counting Task on Implicit Learning in the Serial Reaction Task*, Psychological Research, 61, 83, doi: 10.1007/s004260050015

Frensch, P. A., & Miner, C. S. *1994*, *Effects of Presentation Rate and Individual Differences in Short-Term Memory Capacity on an Indirect Measure of Serial Learning*, 22, 95, doi: 10.3758/BF03202765

Friston, K. *2005*, *A Theory of Cortical Responses.*, Philosophical Transactions of the Royal Society B, 360, 815

Garvert, M. M., Dolan, R. J., & Behrens, T. E. *2017*, *A Map of Abstract Relational Knowledge in the Human Hippocampal–Entorhinal Cortex*, eLife, 6, e17086, doi: 10.7554/eLife.17086

Hadsell, R., Rao, D., Rusu, A. A., & Pascanu, R. *2020*, *Embracing Change: Continual Learning in Deep Neural Networks*, Trends in Cognitive Sciences, 24, 1028, doi: 10.1016/j.tics.2020.09.004

Harrison, S. A., & Tong, F. *2009*, *Decoding reveals the contents of visual working memory in early visual areas*, Nature, 458, 632

He, T., Boudewyn, M. A., Kiat, J. E., Sagae, K., & Luck, S. J. *2022*, *Neural correlates of word representation vectors in natural language processing models: Evidence from representational similarity analysis of event-related brain potentials*, Psychophysiology

Herd, S. A., Banich, M. T., & O'Reilly, R. C. *2006*, *Neural Mechanisms of Cognitive Control: An Integrative Model of Stroop Task Performance and fMRI Data.*, Journal of Cognitive Neuroscience, 18, 22

Herd, S. A., O'Reilly, R. C., Hazy, T. E., et al. *2014*, *A Neural Network Model of Individual Differences in Task Switching Abilities*, Neuropsychologia, 62, 375, doi: 10.1016/j.neuropsychologia.2014.04.014

Hughes, S. W., Lorincz, M., Cope, D. W., et al. *2004*, *Synchronized Oscillations at Alpha and Theta Frequencies in the Lateral Geniculate Nucleus.*, Neuron, 42, 253

Hyafil, A., Summerfield, C., & Koechlin, E. *2009*, *Two Mechanisms for Task Switching in the Prefrontal Cortex*, Journal of Neuroscience, 29, 5135, doi: 10.1523/JNEUROSCI.2828-08.2009

Jung, T.-P., Makeig, S., Westerfield, M., et al. *2000*, *Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects*, Clinical Neurophysiology, 111, 1745

Kahneman, D. 2011, Thinking, Fast and Slow (New York: Farrar, Straus and Giroux)

King, J.-R., & Dehaene, S. *2014*, *Characterizing the dynamics of mental representations: the temporal generalization method*, Trends in cognitive sciences, 18, 203

King, J.-R., Gramfort, A., Schurger, A., Naccache, L., & Dehaene, S. *2014*, *Two distinct dynamic modes subtend the detection of unexpected sounds*, PloS one, 9, e85791

Kingma, D. P., & Ba, J. *2015*, Adam: A Method for Stochastic Optimization, in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, ed. Y. Bengio & Y. LeCun

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al. *2017*, *Overcoming Catastrophic Forgetting in Neural Networks*, Proceedings of the National Academy of Sciences, 114, 3521, doi: 10.1073/pnas.1611835114

Knudsen, E. B., & Wallis, J. D. *2021*, *Hippocampal Neurons Construct a Map of an Abstract Value Space*, Cell, 184, 4640, doi: 10.1016/j.cell.2021.07.010

Koster, R., Chadwick, M. J., Chen, Y., et al. *2018*, *Big-Loop Recurrence within the Hippocampal System Supports Integration of Information across Episodes*, Neuron, 99, 1342, doi: 10.1016/j.neuron.2018.08.009

Kriegeskorte, N., Mur, M., & Bandettini, P. *2008*, *Representational Similarity Analysis - Connecting the Branches of Systems Neuroscience*, Frontiers in Systems Neuroscience, 2

Kumaran, D., & McClelland, J. L. *2012*, *Generalization through the Recurrent Interaction of Episodic Memories: A Model of the Hippocampal System*, Psychological Review, 119, 573, doi: 10.1037/a0028681

Lake, B. M. *2019*, *Compositional Generalization through Meta Sequence-to-Sequence Learning*, arXiv:1906.05381 [cs]. http://ascl.net/1906.05381

Lake, B. M., & Baroni, M. *2018*, Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks, in Proceedings of Machine Learning Research, Vol. 80, Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, ed. J. G. Dy & A. Krause (PMLR), 2879–2888

Lake, B. M., Linzen, T., & Baroni, M. *2019*, Human Few-Shot Learning of Compositional Instructions, in Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada, July 24-27, 2019, ed. A. K. Goel, C. M. Seifert, & C. Freksa (cognitivesciencesociety.org), 611–617

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. *2017*, *Building Machines That Learn and Think like People*, The Behavioral and Brain Sciences, 40, e253, doi: 10.1017/S0140525X16001837

LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. *2013*, *Decoding attended information in short-term memory: an EEG study*, Journal of cognitive neuroscience, 25, 127

Lopez-Calderon, J., & Luck, S. J. *2014*, *ERPLAB: an open-source toolbox for the analysis of event-related potentials*, Frontiers in human neuroscience, 8, 213

Luck, S. J. 2022, Applied Event-Related Potential Data Analysis (LibreTexts), doi: 10.18115/D5QG92. https://doi.org/10.18115/D5QG92

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. *2013*, *Context-Dependent Computation by Recurrent Dynamics in Prefrontal Cortex*, Nature, 503, 78, doi: 10.1038/nature12742

Mathewson, K., Gratton, G., Fabiani, M., Beck, D., & Ro, T. *2009*, *To See or Not to See: Prestimulus Alpha Phase Predicts Visual Awareness.*, 29, 2725

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, 419

McCloskey, M., & Cohen, N. J., Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. 1989, in The Psychology of Learning and Motivation, Vol. 24, ed. G. H. Bower (San Diego, CA: Academic Press), 109–164

Miall, R. C., & Wolpert, D. M. *1996*, *Forward Models for Physiological Motor Control.*, Neural Netw, 9, 1265

Miller, E. K., & Cohen, J. D. *2001*, *An Integrative Theory of Prefrontal Cortex Function.*, Annual Review of Neuroscience, 24, 167

Min, B.-K., & Park, H.-J. *2010*, *Task-Related Modulation of Anterior Theta and Posterior Alpha EEG Reflects Top-down Preparation*, 11, 79, doi: 10.1186/1471-2202-11-79

Mnih, V., Kavukcuoglu, K., Silver, D., et al. *2013*, *Playing Atari with Deep Reinforcement Learning*, arXiv:1312.5602 [cs]. http://ascl.net/1312.5602

Moser, E. I., Kropff, E., & Moser, M.-B. *2008*, *Place Cells, Grid Cells, and the Brain's Spatial Representation System*, Annual Review of Neuroscience, 31, 69, doi: 10.1146/annurev.neuro.31.061307.090723

Mumford, D. *1992*, *On the Computational Architecture of the Neocortex. II. The Role of Cortico-Cortical Loops.*, Biological Cybernetics, 66, 241

Musslick, S., Saxe, A. M., Dey, B., Henselman, G., & Cohen, J. D. *2017*, Multitasking Capability Versus Learning Efficiency in Neural Network Architectures, in Proceedings for the 39th Annual Meeting of the Cognitive Science Society, London, UK, 6

Nili, H., Wingfield, C., Walther, A., et al. *2014*, *A toolbox for representational similarity analysis*, PLoS computational biology, 10, e1003553

Nobre, A., Correa, A., & Coull, J. *2007*, *The Hazards of Time*, 17, 465, doi: 10.1016/j.conb.2007.07.006

Noh, S. M., Yan, V. X., Bjork, R. A., & Maddox, W. T. *2016*, *Optimal Sequencing during Category Learning: Testing a Dual-Learning Systems Perspective*, Cognition, 155, 23, doi: 10.1016/j.cognition.2016.06.007

O'Keefe, J., & Nadel, L. 1978, The Hippocampus as a Cognitive Map (Oxford, England: Oxford University Press)

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. *2011*, *Complementary Learning Systems.*, Cognitive Science, Epub ahead of print

O'Reilly, R. C., & Frank, M. J. *2006*, *Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia.*, Neural Computation, 18, 283

O'Reilly, R. C., Nair, A., Russin, J. L., & Herd, S. A. *2020*, *How Sequential Interactive Processing Within Frontostriatal Loops Supports a Continuum of Habitual to Controlled Processing*, Frontiers in Psychology, 11, doi: 10.3389/fpsyg.2020.00380

O'Reilly, R. C., Ranganath, C., & Russin, J. L. *2021a*, *The Structure of Systematicity in the Brain*, arXiv:2108.03387 [q-bio]. http://ascl.net/2108.03387

O'Reilly, R. C., & Rudy, J. W. *2001*, *Conjunctive Representations in Learning and Memory: Principles of Cortical and Hippocampal Function*, Psychological Review, 108, 311, doi: 10.1037/0033-295x.108.2.311

O'Reilly, R. C., Russin, J. L., Zolfaghar, M., & Rohrlich, J. *2021b*, *Deep Predictive Learning in Neocortex and Pulvinar*, Journal of Cognitive Neuroscience, 33, 1158, doi: 10.1162/jocn_a_01708

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. J. *2013*, *Recurrent Processing during Object Recognition.*, Frontiers in Psychology, 4

O'Reilly, R. C., Wyatte, D., & Rohrlich, J. *2014*, *Learning Through Time in the Thalamocortical Loops*, arxiv. https://arxiv.org/abs/1407.3432

O'Reilly, R. C., Wyatte, D. R., & Rohrlich, J. *2017*, *Deep Predictive Learning: A Comprehensive Model of Three Visual Streams*, arXiv:1709.04654 [q-bio]. http://ascl.net/1709.04654

Panichello, M. F., & Buschman, T. J. *2021*, *Shared mechanisms underlie the control of working memory and attention*, Nature, 592, 601

Park, S. A., Miller, D. S., & Boorman, E. D. *2021*, *Inferences on a Multidimensional Social Hierarchy Use a Grid-like Code*, bioRxiv, 2020.05.29.124651, doi: 10.1101/2020.05.29.124651

Park, S. A., Miller, D. S., Nili, H., Ranganath, C., & Boorman, E. D. *2020*, *Map Making: Constructing, Combining, and Inferring on Abstract Cognitive Maps*, Neuron, 107, 1226, doi: 10.1016/j.neuron.2020.06.030

Park A., S., Zolfaghar, M., Russin, J., O'Reilly, R. C., & Boorman, E. D. *In-prep*, *Representations of the Task Structure in the Brain Account for Cognitive Controls during Decision Making Process*

Pelli, D. G. *1997*, *The VideoToolbox Software for Visual Psychophysics: Transforming Numbers into Movies.*, Spatial Vision, 10, 437

Peters, M., Neumann, M., Iyyer, M., et al. *2018*, Deep Contextualized Word Representations, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)

Radford, A., Wu, J., Child, R., et al. *2019*, *Language models are unsupervised multitask learners*, OpenAI blog, 1, 9

Rao, R. P. *1999*, *An Optimal Estimation Approach to Visual Perception and Learning.*, Vision research, 39, 1963

Rihs, T. A., Michel, C. M., & Thut, G. *2007*, *Mechanisms of selective inhibition in visual spatial attention are indexed by α-band EEG synchronization*, European Journal of Neuroscience, 25, 603

Rose, N. S., LaRocque, J. J., Riggall, A. C., et al. *2016*, *Reactivation of latent working memories with transcranial magnetic stimulation*, Science, 354, 1136

Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. *2005*, *Prefrontal Cortex and the Flexibility of Cognitive Control: Rules Without Symbols.*, Proceedings of the National Academy of Sciences, 102, 7338

Russin, J., O'Reilly, R. C., & Bengio, Y. *2020*, Deep Learning Needs a Prefrontal Cortex, in Bridging AI and Cognitive Science (BAICS) Workshop, ICLR 2020, 11

Russin, J., Zolfaghar, M., Park, S. A., Boorman, E., & O'Reilly, R. C. *2021*, Complementary Structure-Learning Neural Networks for Relational Reasoning, in Proceedings for the 43rd Annual Meeting of the Cognitive Science Society

Russin, J., Zolfaghar, M., Park, S. A., Boorman, E., & O'Reilly, R. C. *2022*, A Neural Network Model of Continual Learning with Cognitive Control, in Proceedings for the 44th Annual Meeting of the Cognitive Science Society

Samaha, J., Bauer, P., Cimaroli, S., & Postle, B. R. *2015*, *Top-down Control of the Phase of Alpha-Band Oscillations as a Mechanism for Temporal Prediction*, 112, 8439, doi: 10.1073/pnas.1503686112

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. *2016*, Meta-Learning with Memory-Augmented Neural Networks

Seabold, S., & Perktold, J. *2010*, Statsmodels: Econometric and statistical modeling with python, in Statsmodels: Econometric and statistical modeling with python

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. *2009*, *Stimulus-specific delay activity in human primary visual cortex*, Psychological science, 20, 207

Shanks, D. R., Implicit Learning. 2005, in Handbook of Cognition, ed. K. Lamberts & R. Goldstone (London: Sage), 202–220

Shenhav, A., Botvinick, M. M., & Cohen, J. D. *2013*, *The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function*, Neuron, 79, 217, doi: 10.1016/j.neuron.2013.07.007

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. *2017*, *The Hippocampus as a Predictive Map*, Nature Neuroscience, 20, 1643, doi: 10.1038/nn.4650

Strohminger, N., Gray, K., Chituc, V., et al. *2016*, *The MR2: A Multi-Racial, Mega-Resolution Database of Facial Stimuli*, Behavior Research Methods, 48, 1197, doi: 10.3758/s13428-015-0641-9

Stroop, J. R. *1935*, *Studies of Interference in Serial Verbal Reactions.*, Journal of Experimental Psychology, 18, 643

Summerfield, C., & de Lange, F. P. *2014*, *Expectation in Perceptual Decision Making: Neural and Computational Mechanisms*, Nature Reviews Neuroscience, 15, 745, doi: 10.1038/nrn3838

Summerfield, C., Luyckx, F., & Sheahan, H. *2020*, *Structure Learning and the Posterior Parietal Cortex*, Progress in Neurobiology, 184, 101717, doi: 10.1016/j.pneurobio.2019.101717

Takagi, Y., Hunt, L. T., Woolrich, M. W., Behrens, T. E., & Klein-Flügge, M. C. *2021*, *Adapting Non-Invasive Human Recordings along Multiple Task-Axes Shows Unfolding of Spontaneous and over-Trained Choice*, eLife, 10, e60988, doi: 10.7554/eLife.60988

Tanner, D., Morgan-Short, K., & Luck, S. J. *2015*, *How Inappropriate High-Pass Filters Can Produce Artifactual Effects and Incorrect Conclusions in ERP Studies of Language and Cognition*, Psychophysiology, 52, 997–1009, doi: 10.1111/psyp.12437

Tanner, D., Norton, J. J. S., Morgan-Short, K., & Luck, S. J. *2016*, *On High-Pass Filter Artifacts (They're Real) and Baseline Correction (It's a Good Idea) in ERP/ERMF Analysis*, Journal of Neuroscience Methods, doi: 10.1016/j.jneumeth.2016.01.002

Tolman, E. *1948*, *Cognitive Maps in Rats and Men.*, Psychological Review, 55, 189

Toosi, T., Tousi, E. K., & Esteky, H. *2017*, *Learning Temporal Context Shapes the Prestimulus Alpha Oscillations and Improves the Visual Discrimination Performance*, jn.00969.2016, doi: 10.1152/jn.00969.2016

Tsuda, B., Tye, K. M., Siegelmann, H. T., & Sejnowski, T. J. *2020*, *A Modeling Framework for Adaptive Lifelong Learning with Transfer and Savings through Gating in the Prefrontal Cortex*, Proceedings of the National Academy of Sciences of the United States of America, 117, 29872, doi: 10.1073/pnas.2009591117

van Ede, F., Niklaus, M., & Nobre, A. C. *2017*, *Temporal expectations guide dynamic prioritization in visual working memory through attenuated α oscillations*, Journal of Neuroscience, 37, 437

VanRullen, R., & Koch, C. *2003*, *Is Perception Discrete or Continuous?*, 7, 207

Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., et al. *1997*, *Differential Effects of Early Hippocampal Pathology on Episodic and Semantic Memory*, Science, 277, 376, doi: 10.1126/science.277.5324.376

Velez, R., & Clune, J. *2017*, *Diffusion-Based Neuromodulation Can Eliminate Catastrophic Forgetting in Simple Neural Networks*, PLOS ONE, 12, e0187736, doi: 10.1371/journal.pone.0187736

Vikbladh, O. M., Meager, M. R., King, J., et al. *2019*, *Hippocampal Contributions to Model-Based Planning and Spatial Memory*, Neuron, 102, 683, doi: 10.1016/j.neuron.2019.02.014

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., et al. *2018*, *Prefrontal Cortex as a Meta-Reinforcement Learning System*, Nature Neuroscience, 21, 860, doi: 10.1038/s41593-018-0147-8

Whittington, J. C. R., Muller, T. H., Mark, S., et al. *2020*, *The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation*, Cell, 183, 1249, doi: 10.1016/j.cell.2020.10.024

Widmann, A., Schröger, E., & Maess, B. *2015*, *Digital Filter Design for Electrophysiological Data – a Practical Approach*, Journal of Neuroscience Methods, 250, 34, doi: 10.1016/j.jneumeth.2014.08.002

Wilcox, E. G., Futrell, R., & Levy, R. *2022*, *Using Computational Models to Test Syntactic Learnability*, Linguistic Inquiry, 1, doi: 10.1162/ling_a_00491

Willingham, D. B. *1999*, *Implicit motor sequence learning is not purely perceptual*, Memory & cognition, 27, 561

Willingham, D. B., Greenberg, A. R., & Thomas, R. C. *1997*, *Response-to-Stimulus Interval Does Not Affect Implicit Motor Sequence Learning, but Does Affect Performance*, Memory & Cognition, 25, 534

Willingham, D. B., Nissen, M. J., & Bullemer, P. *1989*, *On the development of procedural knowledge.*, Journal of experimental psychology: learning, memory, and cognition, 15, 1047

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. *2017*, *Dynamic hidden states underlying working-memory-guided behavior*, Nature neuroscience, 20, 864

Wulf, G., & Shea, C. H. *2002*, *Principles Derived from the Study of Simple Skills Do Not Generalize to Complex Skill Learning*, Psychonomic Bulletin & Review, 9, 185, doi: 10.3758/BF03196276

Wyatte, D. *2014*, *What Happens next and When "next" Happens: Mechanisms of Spatial and Temporal Prediction*. https://arxiv.org/abs/1407.5328

Yonelinas, A., Ranganath, C., Ekstrom, A., & Wiltgen, B. *2019*, *A Contextual Binding Theory of Episodic Memory: Systems Consolidation Reconsidered*, Nature reviews. Neuroscience, 20, 364, doi: 10.1038/s41583-019-0150-4

Zeithamova, D., Schlichting, M. L., & Preston, A. R. *2012*, *The Hippocampus and Inferential Reasoning: Building Memories to Navigate Future Decisions*, Frontiers in Human Neuroscience, 6, doi: 10.3389/fnhum.2012.00070

Zhang, W., Luck, S., & Kappenman, E. *2021*, WHAT BASELINE CORRECTION INTERVAL IS OPTIMAL FOR ERP DATA ANALYSIS?, in PSYCHOPHYSIOLOGY, Vol. 58, WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, S75–S75

Zolfaghar, M., Chaodan, L., Aaron, S., et al. *In-prep*, *From Past to Future: Exploring the Human Brain's Deep Predictive Learning Mechanisms*

Zolfaghar, M., Russin, J., Park, S. A., Boorman, E. D., & O'Reilly, R. C. *2022*, The Geometry of Map-Like Representations under Dynamic Cognitive Control, in Proceedings for the 44th Annual Meeting of the Cognitive Science Society