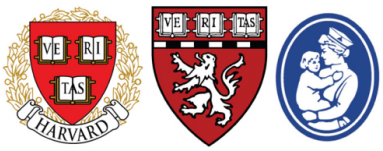
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



Hebbian attractor to model working memory in complex human behavior

Master Thesis

Ravi F. Srinivasan

September 27, 2023

Carried out in the Kreiman Lab, Harvard Medical School
Advisor: Prof. G. Kreiman
Internal advisor: Prof. A. Steger

Department of Computer Science, ETH Zürich

Abstract

Working memory is a crucial cognitive function that enables temporary storage, retrieval, and manipulation of information in goal-directed behavior. While previous research has predominantly dissected these processes in isolation within controlled settings, this study takes a novel approach by examining and modeling the dynamic neural mechanisms underlying multifaceted working memory during a real-world card-matching game. The participants were presented with covered images arranged in a grid and were instructed to flip two images at that time until all matching pairs were found. During this task the participants needed to keep track of the position and content of the images, allowing the analysis of behavioral data and neural activities. We introduced a Hebbian attractor network to model and characterize the memory dynamics of this complex naturalistic task. We showed that the model was able to accurately predict human behavior, and demonstrated similar patterns of memory decay and reaction times. Moreover, we found qualitative equivalents to patients' neural signals that encoded novelty and familiarity, as well as signals that predicted correct retrieval from memory of a tile's pair location. The high temporal resolution, extensive spatial sampling, and computational model provide an opportunity to characterize the dynamics of memory in a complex naturalistic task.

Acknowledgments

First and foremost, I would like to thank Prof. Gabriel Kreiman for providing me with the opportunity to work on this fascinating project. Thank you for the advice during the weekly meetings, the insightful discussions, and for allowing me to connect to different researchers in the field. The continued guidance is very appreciated.

I would also like to thank Dr. Giorgia Dellaferrera and Dr. Martino Sorbaro for introducing me to the field of computational neuroscience and for jumpstarting my research journey. Your impact over the last year has been invaluable. I am thankful to Prof. Angelika Steger for agreeing to oversee this project.

My gratitude goes to all my colleagues in the Kreiman Lab for creating a warm and welcoming environment from the very beginning, for the great advice, for the interesting discussions, and for the fun ones.

I want to extend my heartfelt gratitude to my friends in Milan, Zurich, and Boston for their unwavering encouragement and inspiration, bridging the gap of miles throughout my studies. Lastly, my special thanks go to my family for their unconditional support over the years.

Preface

The work done in this thesis has additionally been submitted as part of a research paper [1].

The task modeled in this manuscript was designed by Yuchen Xiao and Gabriel Kreiman. All the human data used in this work were curated by Yuchen Xiao and analyzed by Yuchen Xiao and Paula Sanchez Lopez with the collaboration of Gabriel Kreiman. The specific contributions are highlighted throughout the text.

At the time of writing, the manuscript has been submitted and is undergoing review. This thesis aims to expand on the paper, giving deeper insights into the computational model.

Abbreviations

STM Short-term Memory

LTM Long-term Memory

WM Working Memory

ML Machine Learning

iEEG Intracranial electroencephalography

ECoG Electrocorticography

sEEG Stereo electroencephalography

NSLC n-since-last-click

NSP n-since-pair

RT Reaction Time

IFPs Intercranial Field Potentials

GLM Generalized Linear Model

AUC Area Under the Curve

VIF Variance Inflation Factor

Contents

List of Figures

Chapter 1

Introduction

Working Memory (WM) plays a pivotal role in our cognitive architecture, enabling us to temporarily store and retrieve immediate information. Unlike many contemporary Machine Learning (ML) algorithms that heavily rely on supervised learning paradigms, the formation and retrieval of memories in biological systems occur in a predominantly unsupervised manner. Memories are forgotten and summoned with single or limited exposures to sensory inputs. This remarkable capability hinges on the brain's capacity to assess and distinguish between novelty and familiarity, establish connections between incoming sensory data and preexisting knowledge, amalgamate spatial and temporal cues, and adeptly retrieve pertinent information in response to current task demands. While extensive research has probed neural responses related to individual facets of WM in controlled laboratory tasks, the holistic orchestration of these components in real-life scenarios remains largely uncharted territory.

One fundamental aspect of WM pertains to non-associative recognition memory, which involves the ability to discern whether a stimulus has been previously encountered. The ability to differentiate between novel and familiar stimuli is a crucial precursor to effective memory encoding [2]. Conversely, recognizing an item as familiar is instrumental in facilitating memory retrieval [3]. Numerous studies have examined neural correlates of recognition memory for novelty versus familiarity, with a primary focus on, but not restricted to, medial temporal lobe structures in various species, including rodents, monkeys, and humans [4–15]. These investigations often centered on tasks that featured the presentation of lists of items, such as words, images, or video clips, followed by either item recall or assessments of recognition memory for those items (e.g. [7, 9–13, 16–21]). Both novel and familiar items must be seamlessly assimilated into the repository of prior knowledge through the formation of novel associations.

In addition to recognition memory, another fundamental component of WM

is associative memory. Associative memory encompasses the capacity to establish connections between items or to evaluate the accuracy of such associations [22–31]. This facet of memory has been frequently explored through tasks that require participants to learn pairs of items and subsequently recall one item when presented with the other or to determine the correctness of given associations. While recognition memory and associative memory have traditionally been studied as separate constructs, they are inherently intertwined in real-world memory tasks. The effective operation of associative memory relies on the foundational processes of recognition memory. Understanding the interplay and distinctions among different memory components during natural and intricate behaviors necessitates a comprehensive investigation.

In this thesis, we delve into the neural underpinnings of Short-term Memory (STM) during complex human behavior and present a computational model that aims to elucidate its cognitive mechanisms. Specifically, we focus on the integration of recognition and WM processes and their neural correlates in a real-life memory task.

1.1 Motivation

The exploration of the neural underpinnings of WM through computational models has been a subject of enduring scientific inquiry aiming to elucidate its neural mechanisms and cognitive significance [32]. Models rooted in persistent neuronal activity [33–41] provide foundational insights into short-term memory's neural basis. Recent perspectives, including those involving attractor networks [42–48], have also highlighted the significance of Hebbian synaptic plasticity and short-term depression and facilitation as means to enhance memory encoding [49–53].

As a proof of principle, we introduce a computational model rooted in attractor-based neural networks. This model represents an attempt to bridge the gap between behavioral and neurophysiological observations and the computational underpinnings of WM during complex human behavior. Our model aims to capture the behavioral and neural responses observed during a naturalistic memory task.

1.2 Thesis Objective

The objective of this thesis is to present and analyze a Hebbian attractor-based computational model that simulates WM during complex human behavior. Through this model, we seek to provide a comprehensive understanding of how recognition and associative memory processes interact and influence neural responses during real-world memory tasks.

To achieve this objective, we employ a multifaceted approach that combines behavioral data collection, intracranial field potential recordings from human participants, and computational modeling techniques. By examining neural activity patterns in response to complex memory tasks and integrating these findings into our computational model, we propose a way to elucidate some of the neural mechanisms governing WM in natural settings.

This thesis describes the motivation, development, evaluation, and application of the Hebbian attractor model to simulate and interpret WM behavior in complex human scenarios. Through experimental insights and computational modeling, this work aims to advance the understanding of WM processes in the context of real-life tasks.

1.3 Structure of the Thesis

The remainder of this thesis is organized as follows:

Chapter 2 provides a detailed review of the relevant literature on STM, recognition memory, associative memory, and computational models of memory.

Chapter 3 outlines the methodology employed in this research, including data collection procedures, neural recording techniques, and the development of the Hebbian attractor model.

Chapter 4 presents the results obtained by analyzing the behavioral data and neurophysiological recordings, detailing the comparison between the human experiments results and the one obtained from the model.

Chapter 5 concludes the thesis by summarizing the key findings, highlighting their significance, and offering a reflection on the broader implications of this research.

Chapter 2

Background

This chapter serves as the essential theoretical foundation, enabling the contextualization and comprehension of the subsequent chapters. It commences with a concise introduction to the concept of WM and includes a review of pertinent literature spanning psychology, biology, and neuroscience. Additionally, it provides an overview of both theoretical and computational models of WM. Within this context, we review Hebbian learning and attractor networks and elucidate how these concepts have been leveraged to unravel the dynamics of brain computation.

2.1 Working memory

2.1.1 Long-term, short-term and working memory

The broader landscape of memory research, spanning psychology, biology, and neuroscience, has experienced more than a century of exploration. Since the inception of the term "memory" by Hermann Ebbinghaus in the 1880s, the field has undergone significant evolution. William James [54] introduced the differentiation between primary and secondary memory. This marked the initial steps toward the establishment of contemporary memory taxonomy, which now encompasses recognized categories such as short-term, long-term, and WM (Fig. 2.1). These categories represent enduring frameworks that have guided extensive research endeavors aimed at unraveling the intricacies of the abstract concept known as memory [55–59].

Long-term Memory (LTM) is the repository of vast knowledge and a record of past experiences. It encompasses all the information accumulated over a lifetime. This form of memory is what allows individuals to recall facts, events, and personal experiences from the distant past. For example, the ability to remember historical dates, childhood memories, or learned skills relies on LTM.

STM, as introduced by James [54], can be seen as the immediate, temporarily accessible storage of a limited amount of information [60]. It functions as a mental workspace for holding and processing information needed for immediate tasks. One classic example is remembering a phone number long enough to dial it or recalling a set of directions while navigating through a new city. STM plays a crucial role in various cognitive activities that require temporary retention and manipulation of information.

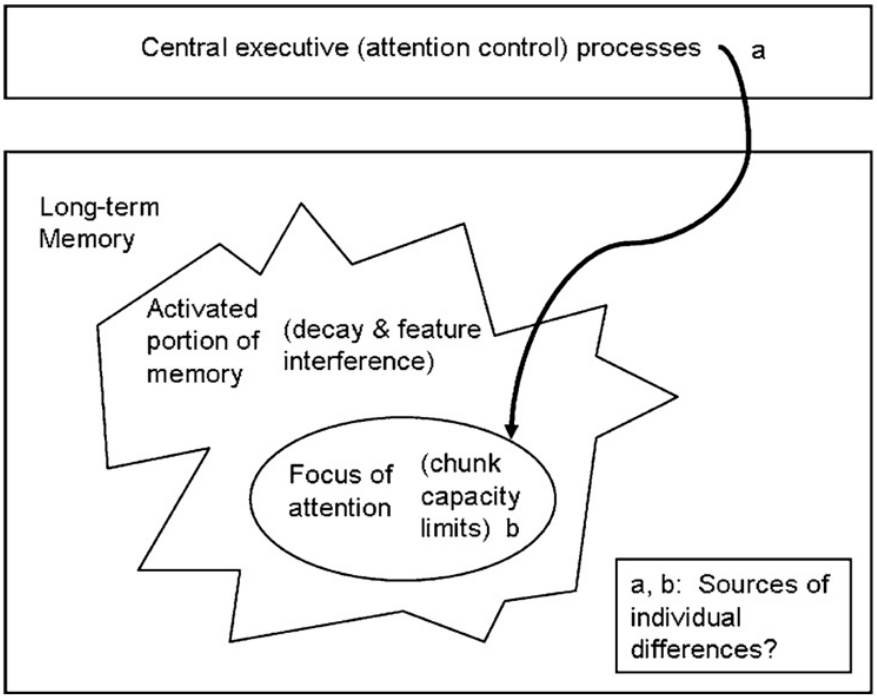


Figure 2.1: A depiction of the theoretical modeling framework. In this framework, STM is derived from a temporarily activated subset of information in LTM. This activated subset may decay as a function of time unless it is refreshed, although the evidence for decay is still tentative at best. A subset of the activated information is the focus of attention, which appears to be limited in chunk capacity (how many separate items can be included at once). New associations between activated elements can form the focus of attention. (Taken from Cowan [61])

WM, a concept refined by Baddeley and Hitch [62], extends beyond STM. It encompasses the active manipulation of information held in STM and its integration into ongoing cognitive processes. WM is involved in complex cognitive tasks that demand both the storage and processing of information simultaneously. For instance, solving mathematical problems, comprehending complex sentences, or following multi-step instructions all rely on WM. It is a dynamic system that helps individuals maintain and manipulate information relevant to their current cognitive endeavors.

2.1.2 Working memory

WM, a critical system involved in the retention of information during complex cognitive tasks such as reasoning, comprehension, and learning, has fascinated scholars since the 1960s [32, 63]. Its influence extends beyond its origins in cognitive psychology, permeating various realms of cognitive science and neuroscience. Moreover, its application has traversed diverse fields, ranging from education and psychiatry to paleoanthropology [32].

The term "WM" made its debut in 1960, attributed to Miller, Galanter, and Pribram in their seminal work, 'Plans and the Structure of Behavior' [64], which conceptualized WM as a form of rapid-access memory utilized in executing plans. Soon after, Pribham et al. [65] speculated that the neural underpinnings of WM involve the prefrontal cortex (PFC), particularly based on observations of deficits resulting from PFC lesions in tasks necessitating delays between stimuli and corresponding responses. Subsequently, in 1968, Atkinson and Shiffrin's influential paper embraced this term [60], which was then adopted by Baddeley and Hitch [62] as the title for a *multi-component* model.

2.1.3 Neural underpinnings of working memory

A seminal study by Fuster and Alexander from 1971 revealed that neurons in the prefrontal cortex exhibited heightened firing rates during delay periods, implying their role in memory maintenance [39]. Subsequent experiments corroborated these findings, further emphasizing the significance of spontaneous neural activity in WM [38, 40, 41, 66]. In fact, inhibiting this persistent activity during delay periods has been shown to diminish task recall accuracy [67]. Moreover, ML models trained on neuronal spike data during this persistent activity have demonstrated the ability to predict animal behavior during recall phases [68]. Consequently, persistent activity within cortical networks has emerged as a leading candidate mechanism for WM, giving rise to computational models such as the ring model, which relies on mutual excitation among neurons to sustain information [34].

Moreover, recent research has unveiled that WM can exist independently of persistent activities, offering a fresh perspective on its neural mechanism. Studies under dual-task conditions have shown that, even when subjects make correct choices, the spatial selectivity of delayed activities diminishes [69]. This implies that WM can be retained within neural states characterized by "activity silence" [70]. Additionally, transient gamma oscillations have been linked to the reactivation of coded sensory information, suggesting their potential role in WM [71]. This has led to the proposal of alternative models where synaptic changes, rather than neural activity, serve as the foundation for information storage. Such models emphasize specific

synaptic mechanisms like presynaptic calcium residue and short-term enhancement [49–51].

While these two mechanisms appear fundamentally distinct, it's essential to recognize that observed experimental phenomena may stem from diverse species, individual conditions, and experimental methodologies, possibly addressing distinct aspects of WM. Moreover, even within research related to persistent activity, only a subset of neurons engaged in the experimental task exhibit persistent activity, displaying diverse temporal patterns [41]. These nuances suggest the potential coexistence of these mechanisms within the brain or an underlying deeper integration. Recent efforts have indeed aimed to unify these seemingly disparate mechanisms into a cohesive framework, offering a new perspective on the neural basis of WM [72].

2.2 Hebbian learning

Hebbian learning, proposed by the Canadian psychologist Donald Hebb in 1949 [73], represents a fundamental concept in neural network theory. At its core, Hebbian learning is a synaptic plasticity rule that posits that "cells that fire together, wire together.". In other words, if two neurons are consistently active at the same time, the strength of the connection between them should increase. This rule can be formally expressed as:

$$\Delta w_{ij} = \eta \cdot x_i \cdot x_j \tag{2.1}$$

Here, Δw_{ij} represents the change in the synaptic weight between neuron i and neuron j, η denotes the learning rate, and x_i and x_j are the activities of neurons i and j, respectively. Importantly, Hebbian learning lacks a mechanism for weakening synapses when neurons do not fire together, which can lead to rapid and uncontrolled network growth. To mitigate this, the learning rule is often refined and combined with other mechanisms, such as synaptic normalization or competition.

Hebbian learning has garnered attention in the neuroscience community due to its biological plausibility. It aligns with observed patterns of synaptic strengthening in biological neural networks. For instance, in the brain's visual cortex, neurons that respond to adjacent regions of the visual field exhibit enhanced synaptic connectivity, reflecting the principle of Hebbian learning. Moreover, experimental studies, including long-term potentiation (LTP) and long-term depression (LTD) in the hippocampus and other brain regions, provide empirical support for the synaptic modifications implied by Hebbian learning.

The concept of Hebbian learning is closely tied to associative memories. This association arises because Hebbian learning strengthens connections between neurons that tend to activate together, effectively encoding correlations in the data. In the context of neural networks, this leads to the formation of attractor states, where specific patterns of neural activity become more stable and attract the network towards particular memory representations. This aligns with the notion of content-addressable memory, where partial or noisy input patterns can retrieve complete stored patterns, a key feature of associative memories. Hebbian learning thus serves as a foundational mechanism for encoding and recalling associations, contributing to our understanding of how the brain forms and retrieves memories [74].

Specifically, Hebbian learning rules enable the establishment of associative connections among neurons that become active together. This allows for the subsequent reactivation of the initial group of active neurons when a partial pattern is reintroduced through associative recall. This short-term Hebbian plasticity, which has been observed in pyramidal neurons, relies on postsynaptic NMDA receptors, manifests rapidly after brief stimulation (e.g., as little as 25 spikes within 500 ms), and can endure for as long as 15 minutes [75]. Lately, various types of rapid-acting Hebbian synaptic plasticity, such as short-term potentiation, have been examined in experiments and put forward as potential contenders for synaptic WM [76, 77]. Short term potentiation becomes evident following brief, high-frequency bursts and notably diminishes in an activity-dependent fashion rather than in a time manner [78].

2.3 Attractor networks

The term "attractor" has gained increasing recognition in neurophysiology as a means to describe stable and stereotyped spatiotemporal dynamics within neural circuits. These dynamics manifest in various forms, such as rhythmic activity in central pattern generators, well-organized propagation patterns of neuronal spike firing in cortical circuits, self-sustained persistent activity in WM processes, and the representation of associative LTM by neuronal ensembles. These intricate neural activity patterns primarily arise through regenerative mechanisms and collective interactions within recurrent networks. The growing interest in attractor networks is driven by the realization that neural circuits often possess numerous feedback loops, and the attractor theory offers a conceptual framework and analytical tools for comprehending these highly recurrent networks.

The concept of attractors originates from the field of dynamical systems mathematics. In a system consisting of interacting units like neurons, given a fixed input, the system typically evolves over time toward a stable state.

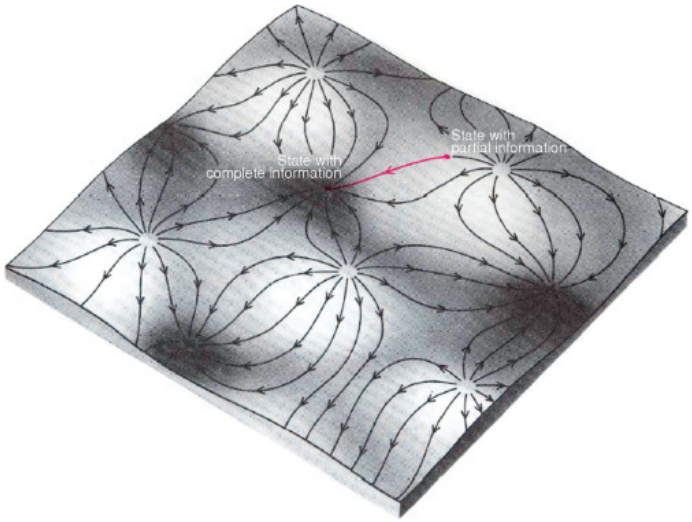


Figure 2.2: Illustration of an attractor model within neural networks. The computational energy function is portrayed as a topographical landscape comprising hills and valleys, plotted against the neural activity states on the XY plane. The configuration of the circuit, including synaptic connections and other intrinsic properties, as well as external inputs, dictate the contours of this landscape. The circuit engages in computation by traversing a trajectory that diminishes the computational energy until it reaches the base of a valley, symbolizing a stable state within the system, often referred to as an attractor. In the context of an associative memory circuit, these valleys correspond to stored memories represented by sets of associated information (the neural activities). When the circuit commences with approximate or incomplete data, it follows a downhill path toward the nearest valley (depicted in red), which contains the complete information. (Taken from Tank and Hopfield [79])

Such stable states are termed attractors because even a minor transient perturbation temporarily alters the system, but it subsequently reconverges to the same state. This concept is visually represented in Fig. 2.2, where a neural network is described by a computational energy function in the space of neural activity patterns. The system's time evolution corresponds to movement downhill in the direction of decreasing computational energy. Each minimum of the energy function represents a stable (attractor) state, while a maximum at the top of a valley represents an unstable state. This depiction can be quantitatively applied to specific neural models.

While characterizing attractor networks as stable and stereotyped may imply insensitivity to external stimuli and a lack of reconfigurability, recent studies have demonstrated the opposite. Attractor networks are responsive to inputs and crucial for the slow-time integration of sensory information in the brain. Moreover, sustained inputs can create or dismantle attractors, allowing the same network to fulfill various functions, such as WM and decision-making, depending on inputs and cognitive control signals. The attractor landscape of a neural circuit can be readily modified by alterations

in cellular and synaptic properties, forming the foundation for the attractor model of associative learning.

2.3.1 Hopfield networks

Hopfield Networks, named after the renowned American physicist John Hopfield, represent a seminal class of recurrent artificial neural networks with associative memory and optimization capabilities [80]. These networks have found significant applications in diverse fields, ranging from pattern recognition to optimization problems.

A Hopfield Network typically consists of a set of binary units or neurons, denoted as s_i, where $i = 1, 2, \ldots, N$, with N being the total number of neurons. Each neuron can take binary values, $s_i \in \{-1, 1\}$, representing an "off" or "on" state. The state of the network evolves over time in discrete steps according to a dynamical update rule. One common update rule is the asynchronous stochastic update, where at each time step, a randomly selected neuron is updated according to:

$$s_i(t+1) = \text{sign}\left(\sum_{j=1}^{N} w_{ij} s_j(t)\right) \tag{2.2}$$

Here, w_{ij} represents the synaptic weight between neuron i and neuron j, which is often symmetric ($w_{ij} = w_{ji}$) and may include self-connections (w_{ii}). The sign$(\cdot)$ function ensures that the state of each neuron remains binary.

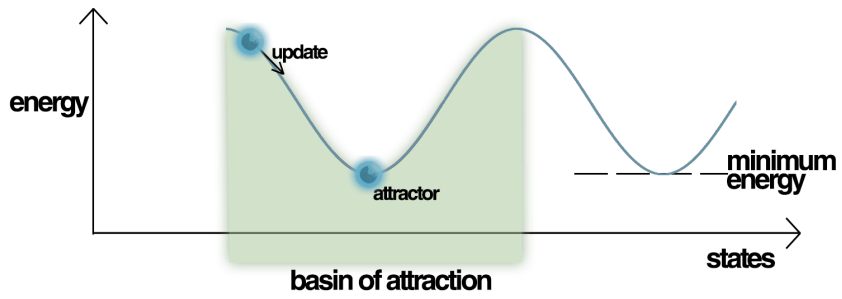


Figure 2.3: Energy Landscape of a Hopfield Network. It depicts the current state of the network (up the hill), an attractor state to which it will eventually converge, a minimum energy level and a basin of attraction shaded in green. Note how the update of the Hopfield Network is always going down in Energy. (Taken from Mrazvan22 [81])

One of the remarkable features of Hopfield Networks is their ability to store and retrieve patterns from their synaptic weights. These networks can serve

as content-addressable memory systems, where given a partial or noisy input pattern, the network can recall and retrieve the complete stored pattern that is most similar to the input. This associative memory property is governed by the network's energy function, known as the Hopfield Energy:

$$E = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}s_i s_j \tag{2.3}$$

The network evolves towards lower energy states (Fig. 2.3), and stable attractor states in the network's dynamics correspond to stored patterns. However, Hopfield Networks have limitations in terms of capacity and noise tolerance.

Continuous Hopfield Networks extend the concept of Hopfield Networks to continuous-valued neurons, often using analog activation functions such as sigmoid or hyperbolic tangent. These networks allow for a broader range of values, not limited to binary states. The dynamics of continuous Hopfield Networks can be described using differential equations, such as:

$$\frac{ds_i}{dt} = -s_i + \sum_{j=1}^{N} w_{ij}\varphi(s_j) \tag{2.4}$$

Where $\varphi(s_j)$ represents the activation function for neuron j. Continuous Hopfield Networks have been used in various applications, including optimization problems and neural modeling, and offer a more flexible representation compared to their binary counterparts.

2.3.2 Representation and memory

Low-dimensional attractor dynamics in neural networks offer a rich array of properties that can significantly impact brain computations. These properties include robust representation of information, memory retention, sequence generation, information integration, and effective decision-making, all of which have been thoroughly examined in scientific literature [82].

Representation and memory form the core of cognitive computation. Representation involves associating inputs with specific states and being able to reliably recall those states when needed. Attractor networks are well-suited for this task as they provide stable internal states that can represent different variables. This mapping from external states to attractor states can be achieved through a learning process.

Memory in attractor networks comes in two forms. First, it's embedded in the network's architecture, defined by the weights that specify the ensemble of attractors. If these weights change based on input, they store long-term information about those inputs. Second, attractor networks can maintain persistent activity within a specific attractor state. When initialized in one of these states, the network tends to stay there for a while, acting as STM for the input that initiated the state. If these memory states can be activated based on their content without specific addressing, they are considered content-addressable, adding complexity to the memory processes in attractor networks.

STM in attractor networks relies on first establishing stable states through long-term synaptic changes. For example, in networks like the Hopfield model, states cannot persist if they were not previously trained to be attractor states. Even STM models based on synaptic facilitation, which are different from persistent activity, implicitly depend on prior long-term plasticity. This is crucial in building neural ensembles that can be reactivated by random inputs. In simpler terms, models without plasticity struggle to explain STM for completely new inputs. However, combining attractors with Hebbian plasticity opens the door to more adaptable STM capabilities.

2.3.3 Working memory

The concept of attractor models extends its influence into the domain of WM [42–45, 47], which pertains to the brain's capacity to temporarily hold information for short durations. Neurons involved in maintaining WM must exhibit sustained activity, even in the absence of direct external input. This persistent activity, a key aspect of WM, is believed to be generated through feedback loop connections. In this context, each neuron in a WM network receives excitatory input from both external sources and intrinsic synaptic connections. Inputs activate specific neuron assemblies, and the resulting spike activity propagates through excitatory synaptic circuits, enabling the network to sustain elevated firing rates when external inputs are removed. Attractor models formalize this concept, proposing that a WM circuit comprises multiple attractor states, each representing a distinct memory item, coexisting with a background or resting state. These attractor states are self-sustained and relatively stable in the face of minor perturbations or noise, yet they can be activated or deactivated by brief external stimuli.

Observations of stimulus-selective neural persistent activity in awake animals performing tasks reliant on WM provide empirical support for the attractor model. For instance, in tasks like the delayed match-to-sample task, where subjects must recall whether two presented visual objects are the same, or tasks involving spatial WM, neurons in various brain regions, including the prefrontal, posterior parietal, inferotemporal, and premotor

cortices, exhibit elevated persistent activity that is selective to specific stimuli. This mnemonic coding encompasses cells tuned to discrete memory items, cells representing spatial information with a bell-shaped tuning function, and cells encoding parametric WM, such as stimulus magnitude, with a monotonic tuning function. These observations align with the attractor model, as they demonstrate that stimulus-selective persistent firing patterns are maintained internally, even in the absence of sensory input, and remain relatively stable over time [83].

However, the heterogeneity and temporal dynamics of mnemonic persistent activity pose challenges to the attractor network model. The specific cellular and circuit mechanisms responsible for generating this persistent activity are still open questions, and researchers are addressing them using biologically constrained models. These models take into account known cortical electrophysiology and emphasize the importance of recurrent connections between neurons, as well as the interplay between excitatory and inhibitory elements, such as the role of N-methyl-D-aspartate (NMDA) receptors in stabilizing WM. Additionally, other processes with time constants of hundreds of milliseconds, such as short-term synaptic facilitation or intrinsic ion channels in single cells, may also contribute to the reverberatory dynamics underlying WM. These complexities further underscore the multifaceted nature of WM and the ongoing quest to unravel its underlying mechanisms.

2.4 Intracranial Electroencephalography

In this manuscript, we compare the model with intracranial field potentials obtained from epileptic patients who underwent implantation of Intracranial electroencephalography (iEEG) electrodes as part of their clinical evaluation. The primary objective of iEEG is to precisely delineate the location of epileptic foci within the brain, a crucial step in the presurgical assessment of epilepsy patients. The ultimate clinical aim is to surgically resect the epileptogenic cortical regions for each individual patient [84]. We analyzed two distinct types of iEEG recordings: (1) Electrocorticography (ECoG), which employs electrodes placed directly on the cortical surface, and (2) Stereo electroencephalography (sEEG), involving wire electrodes that penetrate deeper into the brain tissue [84]. Human iEEG affords the unique capability to capture the activity of a population of neurons with exceptional spatiotemporal precision, thereby providing invaluable insights into the functioning of the human brain.

iEEG records field potentials that encapsulate the aggregate activity of extensive and diverse neural populations. The information conveyed by iEEG signals can be characterized by various frequency bands, including gamma (30-150 Hz), beta (14-30 Hz), and alpha (8-14 Hz) [85]. Specifically, the anal-

ysis conducted by **Yuchen Xiao** and **Paula Sanchez Lopez** focused on the gamma frequency band (30-150 Hz) due to a wealth of research suggesting that high-frequency neuronal activities within this range reflect synchronized firing of neuron ensembles and cortical activations [86, 87]. Existing studies have established that gamma band activities exhibit correlations with neural spiking patterns [71, 88, 89] and have the capacity to convey rich information pertaining to motor control, language processing, memory, and other cognitive functions [71, 90–92].

The literature has documented that different WM modalities, such as STM in our investigation, may be underpinned by the intricate interplay between gamma oscillations and lower-frequency activities, such as those in the beta [71, 93, 94] and alpha bands [95, 96]. A push-pull relationship has been postulated, particularly in the prefrontal cortex: the elevation of beta activity coincides with the suppression of gamma activity, and conversely, leading to the regulation of WM and control over its maintenance [71]. During the encoding and retrieval phases, the default state characterized by beta oscillations is disrupted, with a decrease in beta activities facilitating the increase in gamma activities, thereby enabling the accessibility of information to STM. A similar push-pull dynamic has been reported between the alpha band and neuronal spiking activities [95] in regions encompassing somatosensory and motor cortices, as well as between alpha and gamma power within the visual cortex [96]. While these findings have emerged from studies conducted in non-human primates, our comprehension of how such cross-frequency coupling orchestrates memory processes in humans remains an ongoing area of exploration.

Chapter 3

Methods

In this chapter, we discuss how we approached investigating and addressing the research questions and objectives we outlined in the previous chapters. This will include explaining the task, describing how we collected and analyzed data from human experiments, outlining the model's architecture, and explaining how we collected and analyzed data from the model-based experiments.

3.1 Task paradigm

3.1.1 Human experiments

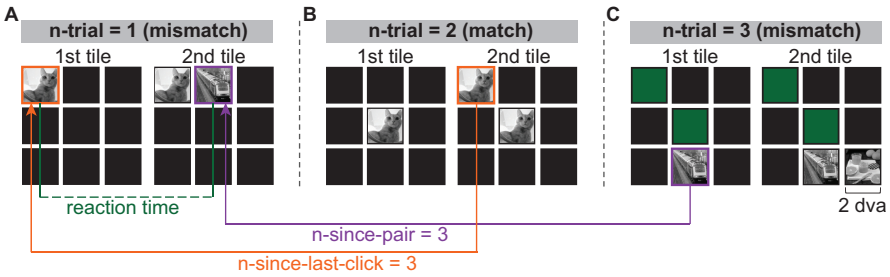


Figure 3.1: Experimental paradigm. A-C. Three consecutive trials in a 3×3 board. In each trial, two tiles were flipped sequentially in a self-paced manner (1st tile, then 2nd tile). If the two tiles contained different images (**A, C, mismatch**), both tiles reset to their original active (black) state after 1 second. If both tiles contained the same image (**B, match**), they turned green after 1 second and stayed green for the remainder of the block. Three behavioral predictors used in the generalized linear models (GLM) are defined here: Reaction Time (RT) (the time between the 1st and 2nd tile within a trial), n-since-last-click (NSLC) (the number of clicks elapsed since the same tile was clicked last), and n-since-pair (NSP) (the number of clicks elapsed since the last time a given tile's matching pair was clicked). Each tile spanned approximately 2 degrees of visual angle (dva) in size. (Figure made by **Yuchen Xiao**)

Participants performed an implementation of the classical memory matching game (Fig. 3.1). The task was designed by **Yuchen Xiao** and **Gabriel Kreiman** and implemented by **Yuchen Xiao**. The game involves remembering the location and content of a set of tiles to find all the matching pairs. A square board containing n × n tiles was shown throughout each block. In the beginning, all tiles were shown in black. In each trial, participants chose one tile, and then a second tile, by clicking on them in a self-paced fashion. Upon clicking, the tile revealed a common object like a cat or an indoor scene like a kitchen. At the end of each trial, either the two tiles revealed the same content (match) or not (mismatch). If the tiles matched, then the two tiles turned green 1,000 ms after the second click, and the two tiles could not be clicked again for the remainder of the block. If the tiles did not match, they turned black 1,000 ms after the second click and could be clicked again in subsequent trials. When all tiles turned green, i.e., all matches were found, the block ended, and another block began. During each block, the map between positions and objects was fixed. The game always started with a block of size 3×3 and progressed to more difficult blocks (4×4, 5×5, 6×6, and finally 7×7). Blocks with an odd number of tiles (3×3, 5×5, and 7×7) contained one distractor object (a human face) with no corresponding pair. For each block except the 3×3 board, there was a limit for the total time elapsed (2 minutes for 4×4, 3.3 min for 5×5, 4.8 min for 6×6, and 8.2 min for 7x7). If a participant did not complete a block within the time limit, the block ended, and a new, easier block started by reducing the board size n by 1, except when n=7, where it was reduced by 2. Conversely, when participants successfully completed a block with a board of size n within the allotted time limit, they moved on to a more difficult block by increasing n by 1. When participants completed an n=7 block, they performed further n=7 blocks. There was no image repetition across blocks.

All the images were from the Microsoft COCO 2017 validation dataset [97] and were rendered in grayscale and square shape. We included a balanced number of pictures from 5 categories: person, animal, food, vehicle, and indoor scenes. All the images were rendered on a 13-inch Apple MacBook Pro laptop. The size of each tile was 0.75×0.75 inches (approximately 2x2 degrees of visual angle, dva) and the separation between two adjacent tiles was 0.125 inch (0.33 dva) for board size n=7 and 0.25 inch (0.67 dva) for the others. The game implementation was written and presented using the Psychtoolbox extension [98, 99] in Matlab_2016b (Mathworks, Natick, MA).

3.1.2 Model experiments

Our computational model undertook a modified version of the task outlined in the preceding paragraph. It tackled this task across 20 boards for each board size, ranging from 3x3 to 7x7.

In the human experiments, participants actively exercised choice in selecting the first tile, a process influenced by elements beyond the scope of the model we sought to construct—namely, strategy and positional bias. Consequently, we directed our model's development towards capturing the number of clicks to solve the task and the behavioral metrics associated with the second tile selection. In the model's execution of the task, the initial tile choice was made uniformly at random from the available tiles, while the model employed information related to the first tile and its own internal state to make its selection for the second tile.

Furthermore, our model operates based on two additional fundamental assumptions. Firstly, we presuppose that patients possess the capacity to perfectly encode the images they are presented with. This presupposition entails that any likeness or similarity between images does not exert an influence on how memories of these similar images might interfere with the task at hand. Secondly, we assume that the grid's positional factors do not exert any influence on the memorization process. This implies that we disregard any potential edge effects and the proximity of errors to the accurate tile location. These assumptions are represented by one-hot encoding both the position and label of the tiles. These assumptions are based on the analysis of the human behavioral data explained in Sections 3.3.2 and 3.3.3.

3.2 The computational model

3.2.1 Model architecture and dynamics

We developed an attractor network model consisting of a fully connected recurrent network with the number of units n equal to the number of tiles in the grid plus the number of different images. For example, the model for the 3x3 board shown in (Fig. 3.2.A) was an attractor network with n=3x3+5=14 units.

The units in the network were designed to model "where" and "what", i.e., position and image labels. Let $\mathbf{x}_p$ be a vector of length equal to the number of tiles in the grid, $\mathbf{x}_l$ be a vector of length equal to the number of different images in the grid, and $\mathbf{x}$ denote the concatenation $[\mathbf{x}_p, \mathbf{x}_l]$ (Fig. 3.2.A). The input to the network is $\mathbf{x}$. Each entry in $\mathbf{x}_p$ and $\mathbf{x}_l$ can take the values -1, 0, or 1. The state of the network at time t is denoted by the vector $\mathbf{h}_t = [\mathbf{p}_t, \mathbf{l}_t]$ of size n, where $\mathbf{p}_t$ and $\mathbf{l}_t$ are the vectors of activations of the position and label units, respectively. Each entry in $\mathbf{h}_t$ is a scalar value. The units in the network are connected in an all-to-all fashion and the matrix $\mathbf{M}_t$ indicates the weights at time t ($\mathbf{M}_t \in \mathbb{R}^{n \times n}$).

The network stores memories in both persistent activities (active representations) and weights (silent representations) [45]. In contrast to Manohar et al.

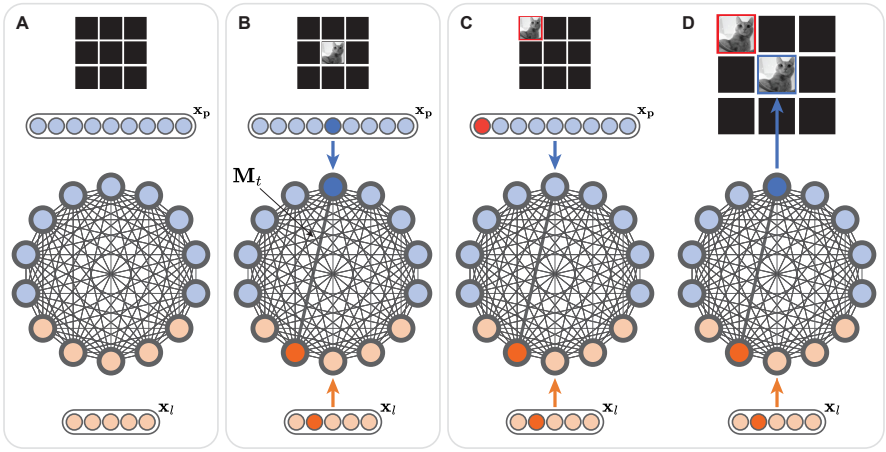


Figure 3.2: Hebbian attractor model architecture and operating regimes. A. Schematic representation of the model architecture used for the 3x3 grid. The 9 blue units encode position ($\mathbf{x}_p$), while the 5 orange units represent the image label ($\mathbf{x}_l$). The black lines between units illustrate the Hebbian weights $\mathbf{M}_t$ in the attractor network. **B.** Learning regime. In this example, the model represents a cat (label=2) at position=5. The dark color indicates strong activation of the corresponding units. **C, D.** Inference regime. In this example, the model is tasked with matching the cat (label=2) observed at position=1. Only the label information is provided to the model in the inference regime. The model's updates (Section 3.2) lead to the unit representing position=5 to exhibit the highest activity (**D**), thereby determining the corresponding tile to be clicked. The darker color indicates stronger activation of the corresponding units. The red color indicates the tile to match (unavailable, Section 3.2) and the corresponding positional unit.

[45] approach, which incorporates a bottleneck in the model to restrict its capacity, our model lacks any such bottleneck. Given an input $\mathbf{x}$ at time t, the network state and weights were updated similarly to Ba et al. [100], i.e., according to:

$$\mathbf{h}_t = f\left(\mathcal{N}\left(\mathbf{x} + \mathbf{M}_{t-1}\mathbf{h}_t\right)\right) \tag{3.1}$$

$$\mathbf{M}_t = \lambda \mathbf{M}_{t-1} + \eta \mathbf{h}_t \mathbf{h}_t^\top \tag{3.2}$$

Here $f(\cdot)$ is the LeakyReLU activation function and $\mathcal{N}(\cdot)$ is activation normalization. λ and η represent a decay rate for the previously stored memories and the learning rate for new memories, respectively. Before the start of each board, the network weights were initialized uniformly at random in $[0, 1]$, while the state of the network was initialized to 0. We note that the Hebbian learning is computed on the state of the network $\mathbf{h}_t$ rather than on the input $\mathbf{x}$. This means that the update of the memory matrix $\mathbf{M}_t$ is influenced by the interference between active and silent representations, thus limiting the network capacity.

3.2.2 Operation regimes

The model operates in two distinct regimes, which we refer to as *learning* (Fig. 3.2.B) and *inference* (Fig. 3.2.C-D). For each trial, the 1st tile was chosen at random among the available tiles. To simulate the gameplay, for each trial, the model performs learning→inference→learning. First, the model learns the position and label of the 1st tile. Second, the model performs inference on the label of the 1st tile. At the end of the inference regime, the most active neuron in determines which tile to click (Fig. 3.2.D). Last, the model learns the position and label of the 2nd tile.

During learning (Fig. 3.2.B), the corresponding position entry of $\mathbf{x}_p$ is set to 1 and all other units are set to -1. Similarly, the corresponding label entry of $\mathbf{x}_l$ is set to 1 and all other units are set to -1. The network dynamics goes through 10 steps according to Eqs. (3.1) and (3.2). During inference (Fig. 3.2.C-D), the corresponding label $\mathbf{x}_l$ of is set to 1 and all the other units are set to -1. All the units of $\mathbf{x}_p$ corresponding to the available tiles are set to 0, while the ones corresponding to the unavailable tiles are set to -1. The network dynamics goes through 10 steps according to Eqs. (3.1) and (3.2). After these 10 steps, we select the unit with the maximum activation within the units of $\mathbf{x}_p$ corresponding to available tiles. If the second tile is a match, then those two tiles become unavailable in the next trials. The weight matrix $\mathbf{M}_t$, however, continues to include all the connections among all the units. The model proceeds until all tiles have been matched. The algorithms for implementing these regimes are detailed in Appendix A.

3.3 Behavioral analysis

3.3.1 Human data

Two computational models were created to simulate behavior assuming perfect memory or no memory (chance performance, Fig. 4.1). The perfect memory model remembered all revealed tiles without forgetting. The random model simulated random clicking. The analyzed data included the reaction time (RT, time between two clicks in a trial), n-since-pair (NSP, number of clicks since the last time a tile's matching pair was seen), n-since-last-click (NSLC, the number of clicks since the same tile was clicked). For NSP and NSLC, the trials in which any tile was seen for the first time were excluded, i.e., when a tile's matching pair had never been revealed, or there was no previous click. These variables were compared for match and mismatch trials at each board size (Figs. 4.3 and 4.4, permutation test, 5,000 iterations, α=0.01). Random matches were defined as a match trial where the second tile had never been seen before; such trials were excluded from both the behavioral and neurophysiological analyses. The F-test was used for linear regression models to assess whether RT, NSP, NSLC, and n-times-seen

significantly covary with board size. The linear regression models' predictors were these four behavioral parameters and the dependent variable the board size. We created separate models for match and mismatch trials and 1st and 2nd tiles. The human behavioral data analysis was conducted by **Yuchen Xiao** and **Paula Sanchez Lopez**.

3.3.2 Random-perfect and perfect memory

We conducted a comparison between two models: the perfect memory model and the random-perfect memory model. In the perfect memory model, each tile's position and label are stored whenever it's clicked. Before each trial, this model scans its memory to identify if there are two tiles with matching labels in distinct positions. If such a pair is found, these two tiles are selected for the trial. In contrast, the random-perfect memory model selects the first tile of each trial uniformly at random and then relies on its stored memory to choose the second tile. We observe that there is minimal disparity between these two models in terms of the N clicks per tile metric (Fig. 3.3).

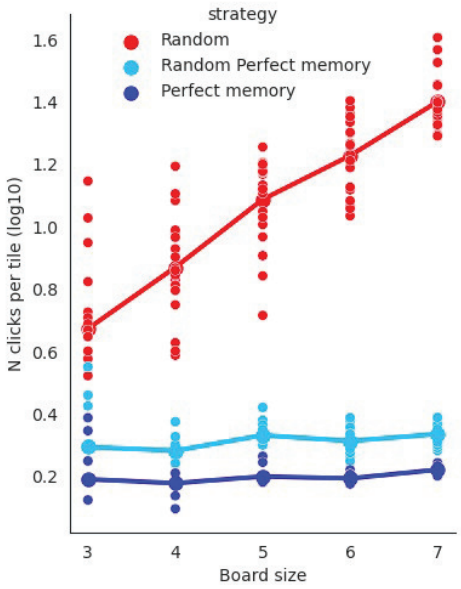


Figure 3.3: N clicks per tile reported for random memory, random-perfect memory, and perfect memory.

3.3.3 Distance from correct tile

We conducted supplementary analyses on the human behavioral data to probe whether, during mismatch trials, participants tend to choose tiles situated near the correct tile. Specifically, for mismatch trials across varying

board sizes, we calculated the Euclidean distance between the correct tile and the tile chosen.

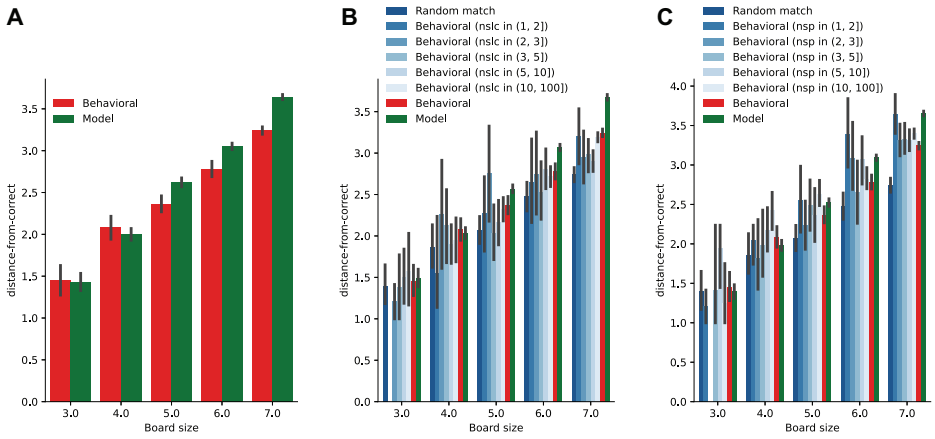


Figure 3.4: Average distance from correct tile in mismatch trials. A-C. Comparison between random model and human behavioral data. **B.** The distance in the human behavioral data is binned by the value of the NSLC. **C.** The distance in the human behavioral data is binned by the value of the NSP.

Our findings reveal a minimal distinction between the behavior of a random model and that of human participants concerning the distance from the correct tile (Fig. 3.4.A). Furthermore, we observed that this distance is not affected by how recently the same tile was last clicked (Fig. 3.4.B) or by the last time a matching tile was selected (Fig. 3.4.C).

We extended our analysis to examine the positions of selected tiles in both match and mismatch trials, to find potential edge effects. We computed the frequency of match (Fig. B.1) and mismatch trials (Fig. B.2) for each position in the board for each board size, as well as the difference between frequencies of matches and mismatches (Fig. B.3). We conducted an identical experiment using a random memory model (20 trials per board size). It is worth noting that our investigation did not yield any results indicating the presence of corner or edge effects. However, the visual representations in our analysis suggest the possibility of the impact linked to participants' bias in selecting tiles during the first trials and potential primacy effects in memorization. Nevertheless, further analysis is required to confirm these observations definitively.

3.3.4 Model data

The N clicks per tile, NSLC, and NSP click for the 2nd tile were calculated for the model identically as for the humans and compared to the participant

behavior (Figs. 4.3 and 4.4).

To compute a proxy for the RT in the model, we used the same approach as in Manohar et al. [45], whereby the unit in $\mathbf{x}_p$ with the strongest activation during the inference time was selected and the RT was computed as the number of steps the unit takes to reach 0.9 of its maximum value.

3.4 Neurophysiological recordings

3.4.1 Epilepsy participants and recording procedures

Intercranial Field Potentials (IFPs) from 20 patients with pharmacologically intractable epilepsy (12-52 years old, 9 female) were recorded by **Yuchen Xiao** and **Ruijie Wu**. These patients were undergoing monitoring at Boston Children's Hospital (Boston, US), Brigham and Women's Hospital (Boston, US), and Xuanwu Hospital (Beijing, China) were rec. All recording sessions were seizure-free. All patients had normal or corrected-to-normal vision. The study protocol was approved by each hospital's institutional review board. Experiments were run under patients' or their legal guardians' informed consent. One patient at Brigham and Women's Hospital (BWH) was implanted with both sEEG and ECoG electrodes, while all other patients had only sEEG electrodes (Ad-tech, USA; ALCIS, France). IFPs were recorded with Natus (Pleasanton, CA) and Micromed (Italy). The sampling rate was 2048 Hz at Boston Children's Hospital, 512 Hz or 1024 Hz at BWH, and 512 Hz at Xuanwu Hospital. Electrode trajectories were determined based on clinical purposes for precisely localizing suspected epileptogenic foci and surgically treating epilepsy [101].

3.4.2 Electrode localization

Electrodes were localized using the iELVis [103] toolbox. **Yuchen Xiao** used Freesurfer [104] to segment the preimplant magnetic resonance (MR) images, upon which post-implant CT was rigidly registered. Electrodes were marked in the CT aligned to preimplant MRI using Bioimage Suite [105]. Each electrode was assigned to an anatomical location using the Desikan-Killiany [102] atlas for subdural grids or strips or FreeSurfer's volumetric brain segmentation for depth electrodes. For white matter electrodes, we also reported their closest gray matter locations. Out of 1,750 electrodes in total, 676 bipolarly referenced electrodes in the gray matter were included (Fig. 3.5) and 492 bipolarly referenced electrodes in the white matter. Five hundred eighty-two electrodes were not considered for analyses due to bipolar referencing, locations in pathological sites, or electrodes containing large artifacts. Electrode locations were mapped onto the MNI305 average brain via affine transformation [106] for display purposes (e.g., Fig. 3.5).

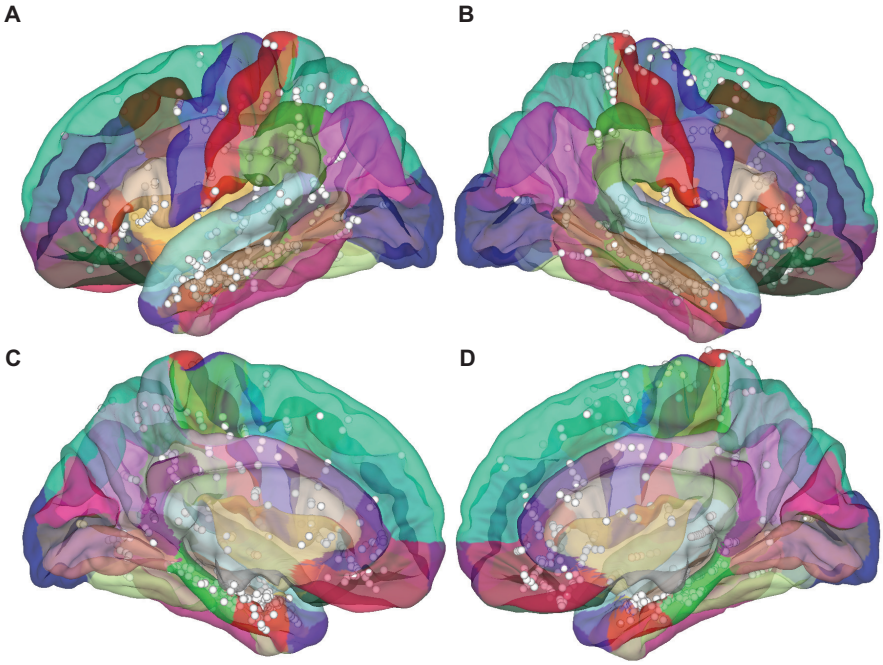


Figure 3.5: Locations of electrodes in the gray matter. Each circle shows a bipolarly referenced electrode (n=676), overlayed on the Desikan-Killiany Atlas with different views: **A.** left **D.** right medial. The colors reflect the Desikan-Killiany parcellation [102]. (Figure made by **Yuchen Xiao**)

3.4.3 Preprocessing of intracranial field potential data

Bipolar subtraction was applied to each pair of neighboring electrodes on each shank of depth electrodes or subdural grids/strips [107]. A zero-phase digital notch filter (Matlab function "filtfilt") was applied to the bipolarly subtracted broadband signals to remove the line frequency at 60 Hz (BCH, BWH) or 50 Hz (Xuanwu) and their harmonics. For each electrode, trials whose amplitudes (Voltagemax-Voltagemin) were larger than 5 standard deviations from the mean amplitude across all trials were considered potential artifacts and discarded from further analyses [108]. For the first tile, the time window for artifact rejection was from 400 ms before the click until 1 second after the average RT. For the second tile, the time window was [400 ms + average RT] before the second click until 1 second after the second click. Across all electrodes, 1.75% of all trials for the 1st tile and 1.73% for the second tile we rejected. These analyses were conducted by **Yuchen Xiao** and **Paula Sanchez Lopez**.

3.4.4 Time-frequency decomposition

The gamma band (30-150 Hz) power was computed using the Chronux toolbox [109]. A time-bandwidth product of 5 and 7 leading tapers, a moving window size of 200 ms, and 569 a step size of 10 ms [110] were used. For each trial, the power was normalized by subtracting the mean gamma band power during the baseline (400 ms before 1st tile) and dividing by the standard deviation of the gamma power during the baseline. For all the participants, there were more mismatches than match trials. In the raster plots, we subsampled the mismatch trials, keeping those trials whose RTs were closest to the mean RT of match trials. All random matches were excluded from analyses. These analyses were conducted by **Yuchen Xiao** and **Paula Sanchez Lopez**.

3.4.5 Generalized linear models

Generalized Linear Models (GLMs) [111, 112] were employed to analyze the relationship between gamma-band power and behavioral parameters in this study, focusing on neural responses between the 1st and 2nd tiles. The response variable for the GLM analyses was defined as the Area Under the Curve (AUC) of gamma-band power within the specified time windows. For computing the AUC, the analysis window commenced when the 1st tile was clicked and concluded at a time corresponding to the 90th percentile of the reaction time distribution, a choice made to balance minimizing overlap with responses after the 2nd tile and maximizing the information captured.

Multicollinearity analysis was performed to assess the presence of highly correlated predictors that could impair the model's performance. Variance Inflation Factor (VIF) for each predictor was calculated to detect the presence of multicollinearities. A VIF of 1 indicates that there is no correlation with other predictors. The larger the VIF, the higher the correlation. A VIF greater than 5 indicates a very high correlation that could significantly harm the model's performance. For all participants in this analysis, the VIFs of all predictors were smaller than 3.

For each predictor, the parameter estimate (beta coefficient) were computed based on the least mean squares fit of the model to the data, the t-statistic (beta divided by its standard error), and the p-value to assess the impact of each predictor on the neural responses. A beta coefficient or t-statistic of zero indicated that the predictor had no effect on the neural responses. A predictor was considered statistically significant if the GLM model differed from a constant model ($p<0.01$), and the p-value for that predictor was less than 0.01. These analyses were conducted by **Yuchen Xiao** and **Paula Sanchez Lopez**.

3.4.6 Mapping the model to neurophysiological recordings

In this work, we modeled the response of two selected electrodes (Fig. 3.6). We selected these electrodes based on the interpretable and interesting properties they exhibited. For the selected electrode in the left pars opercularis (Fig. 3.6.A-B), the NSLC was a significant predictor, while for the selected electrode in the right lateral orbirtontal cortex (Fig. 3.6.C-D) weather a trial was a match or a mismatch was a significant predictor. The significance was determined using the GLMs (Section 3.4.5). We refer to the signal encoded by these electrodes as *surprise* (Fig. 3.6.B) and *confidence* (Fig. 3.6.D) signals.

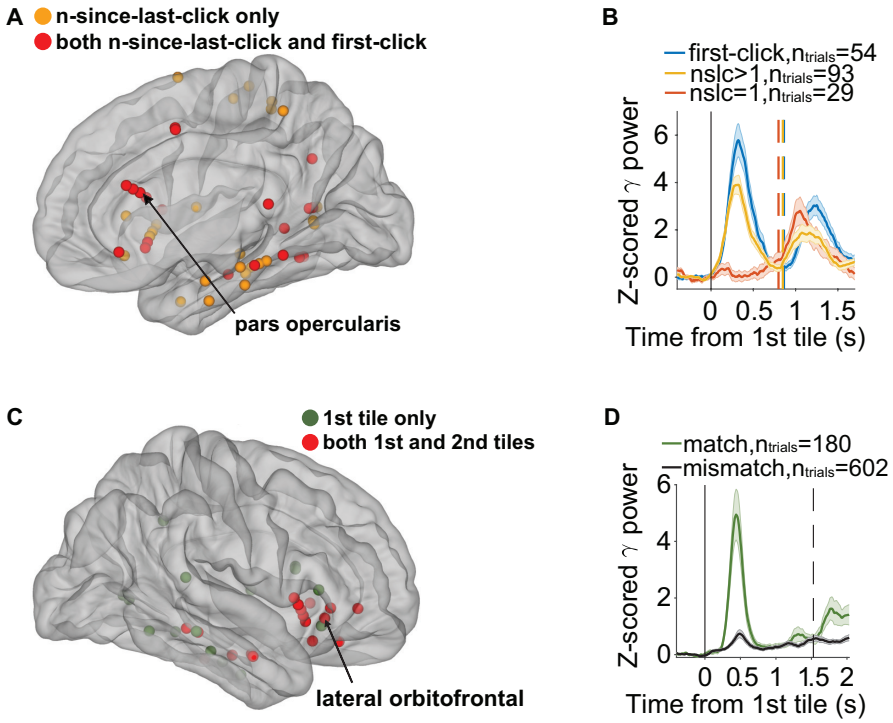


Figure 3.6: The two electrodes which were selected to be modeled using the Hebbian attractor network. A, B. Selected electrode in the left pars opercularis. **A.** Location of the electrode. **B.** Z-scored gamma band power aligned to the 1st tile onset (solid vertical line) for novel tiles (blue), unfamiliar tiles (n-since-last-click>1, yellow), and familiar tiles (n-since-last-click=1, red). The vertical dashed line indicates the mean RT. Multiple dashed lines indicate RT equalization (Section 3.4.3). The time axis extends from 400 ms before the click to 500 ms after the average RT. **C, D.** Selected electrode in the right lateral orbitofrontal cortex. **C.** Location of the electrode. **D.** Z-scored gamma band power aligned to the 1st tile onset (solid vertical line) for match trials (green) and mismatch trials (black). The vertical dashed line indicates the mean RT. Shaded error bars indicate s.e.m.

Novelty/Familiarity signal First, we defined the max-energy metric computed during the 1st learning phase of each trial, in analogy to the memory

signals in Fig. 3.6.A. The energy of the network was computed as:

$$E_t = -\mathbf{h}_t \mathbf{M}_t \mathbf{h}_t^\top \tag{3.3}$$

Min-max normalization was applied to the energy in each trial, and the maximum value in each trial was reported. The model's max-energy signal during gameplay is shown in Fig. 4.5.B.

Confidence signal Second, we defined a confidence metric that reflected the evidence for a match in a given trial, in analogy with the predictive signals shown in Fig. 3.6.B. The confidence metric was defined by selecting the strongest activation in $\mathbf{p}_t$ during inference, subtracting the mean value of $\mathbf{p}_t$, applying min-max normalization to the difference, and then taking the maximum over time t in each trial. The model's confidence signal during gameplay is shown in Fig. 4.6.B.

3.5 Model selection

We tuned the model's hyperparameters to align with the N clicks per tile metric derived from human behavioral data. To accomplish this, we conducted a grid search to find the optimal values for the learning rate (η) and decay rate (γ), exploring the parameter space within the range of (0, 1] with increments of 0.05.

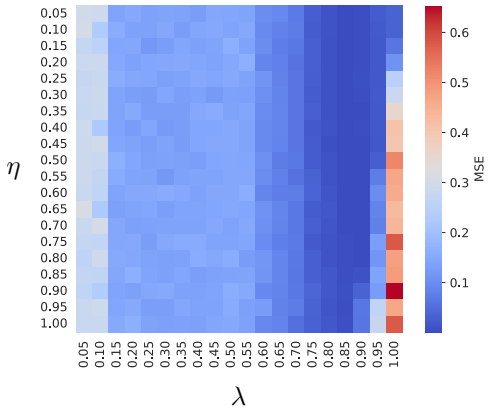


Figure 3.7: Results of the grid-search. Each square color indicates the value of the MSE between human and model N clicks per tile found during the grid-search.

For each combination of these parameter values, for each board size (from 3x3 to 7x7), the model executed the task and recorded the N clicks per tile.

Subsequently, we computed the mean N clicks per tile for both the model's simulations and the participant data for each board size. We then computed the Mean Squared Error (MSE) between these two sets of values. The model that achieved the lowest MSE was the one configured with $\eta = 0.9$ and $\lambda = 0.6$. These parameter settings were then employed consistently across all simulations in this study.

In accordance with Ba et al. [100], the implementation of activation normalization renders the network robust when it comes to selecting suitable decay and learning rates. Our experiments align with this notion as we observe a smooth landscape in the MSE across the explored range of hyperparameter values. Notably, even when we deviate from the optimal hyperparameters, we observe minimal influence on the MSE value (Fig. 3.7).

Chapter 4

Results

We recorded IFPs from 20 patients with pharmacologically intractable epilepsy implanted with depth electrodes (Fig. 3.5). These participants played a memory-matching game (Section 3.1.1, Fig. 3.1). Each trial consisted of two self-paced clicks. Clicking on a tile revealed an image (Fig. 3.1.A). Image categories included person, animal, food, vehicle, and indoor scenes. If the two tiles in a trial contained the same image (match, Fig. 3.1.B), the two tiles turned green and could not be clicked again for the remainder of the block. If the two images were different (mismatch, Fig. 3.1.A, C), the two tiles turned black and could be clicked again. Participants started in a 3×3 tile board block like the one shown in Fig. 3.1 and progressed to more difficult blocks (4×4, 5×5, 6×6, or 7×7 tiles). All tiles had a corresponding match, except for one tile in the boards with an odd number of tiles (3×3, 5×5, and 7×7). We ran the same experiments on the computational model, which was tested for 20 trials for each board size.

4.1 Human experiments

4.1.1 Behavioral data

As board difficulty increased, the average number of clicks per tile also rose as expected (Fig. 4.1.A). Participants outperformed a memoryless model (random clicking) by a significant margin ($p<0.001$, permutation test with 5,000 iterations, one-tailed), but they fell short of a model assuming perfect memory ($p<0.001$, Fig. 4.1.A). RTs, defined as the time between the first and second clicks within a trial, was consistently longer for mismatch trials compared to match trials for all board sizes ($p<0.007$, Fig. 4.2.B).

In each trial, NSLC represented the number of clicks since the last time the same tile was clicked (Fig. 3.1.A-B). For the 2nd tile, NSLC was larger in mismatch trials than in match trials for all board sizes except the 3×3 case

($p<0.001$, Fig. 4.3.A), indicating a memory decay for tiles not recently seen. Another measure, NSP, counted the number of clicks since the last time a tile's matching pair was seen (Fig. 3.1.A, C). For the 2nd tile, NSP was always one in match trials, as the matching pair was revealed in the previous click, resulting in a significant difference between match and mismatch trials ($p<0.001$, Fig. 4.4.A).

4.1.2 Intracranial field potentials

IFPs were recorded from numerous electrodes, with some excluded due to various factors. Results from electrodes in the white matter were also considered in the analysis. A GLM was developed to characterize how neural responses related to cognitive demands in each trial, focusing on the 1st tile.

The GLM considered various predictors, including match status, RTs, NSLC, NSP, and additional factors like first-click, number of times an image had been seen, board size, tile position, and image content. Multicollinearity was addressed using the variance inflation factor, confirming that predictor correlations didn't impact model performance.

Neural responses to the 1st tile were found to correlate with novelty, with decreased activity for novel tiles (Fig. 4.5.A). This correlation was observed in specific brain regions. Familiarity, as indicated by NSLC, also influenced neural responses. Interestingly, novelty and familiarity effects persisted even after reaction time equalization.

Neural signals before the 2nd tile appeared predictive of the trial's outcome (match or mismatch), suggesting that participants internally retrieved pair locations (Fig. 4.6.A). These predictions were observed in specific brain regions, such as the lateral orbitofrontal cortex, medial temporal lobe, and insula.

4.2 Model experiments

We built a computational model that focused on the storage and retrieval of information (Fig. 3.2, Section 3.2). The computational model consists of a Hebbian attractor neural network with all-to-all connectivity. The units are divided into position units (the number equaling the number of tiles on the board) and label units (the number equaling the number of images on the board) (Fig. 3.2.A). The model has two main modes of operation: learning (Fig. 3.2.B), and inference (Fig. 3.2.C-D). After the first click, the model receives as input the label of the tile and its position. The activity of each unit evolves over time based on the input and the weighted input from other units followed by a rectifying non-linearity and normalization (Section 3.2, Eq. (3.1)). Concomitantly, the weights are updated in a Hebbian manner

(Section 3.2, Eq. (3.2)). During inference, the model selects the position unit with the maximum activation for the second click. The model proceeds in this manner until all matches have been found.

4.2.1 The model predicts human behavior

We evaluate the performance of the model using the same evaluators as in the human experiments, i.e. total number of clicks per board size, RT, NSLC, NSP. We did not compute NSLC and NSP for the first tile because the model chooses the first tile randomly among the available tiles (Section 3.1.2). We defined the reaction time as the number of steps needed for the selected unit to reach 0.9 of its maximum value (Section 3.3.4). The model was fit to match the total number of clicks per tile for each board size (Section 3.5).

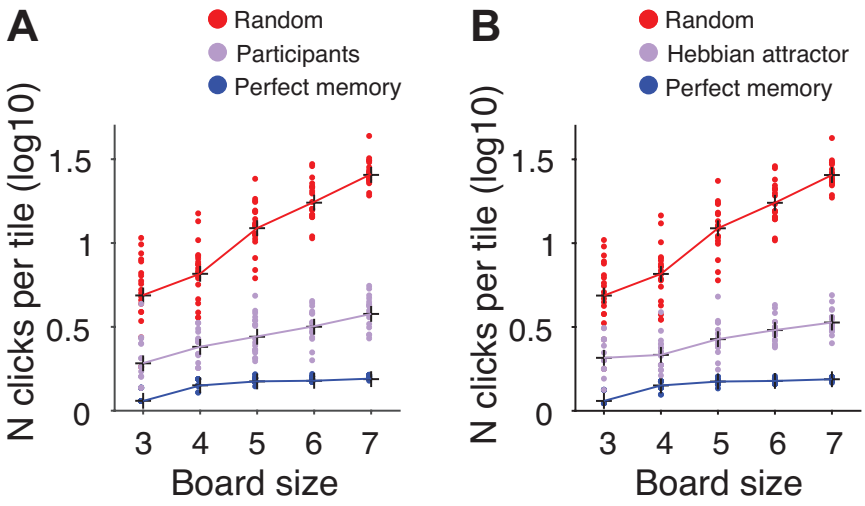


Figure 4.1: Number of clicks per tile (log scale) as a function of board size. Figure shows metric for random simulation model (red, n=20), perfect memory simulation model (blue, n=20), and epilepsy patient participants (**A**, purple, n=20) and Hebbian attractor (**B**, purple, n=20) (Section 3.3). Perfect memory simulation models may generate a different number of clicks per tile because the click location for new tiles was randomized. The performance of both the epilepsy patients and the Hebbian attractor was better than the random model and worse than the perfect model. In both cases, the number of clicks per tile increased as board size increased.

N-click-per-tile Fig. 4.1 illustrates the number of clicks per tile (log scale) as a function of board size for different models, including a random simulation model, a perfect memory simulation model, epilepsy patient participants, and our Hebbian attractor model. Notably, the number of clicks per tile increased with the board size, closely resembling the behavior exhibited by participants (compare Fig. 4.1.A and B).

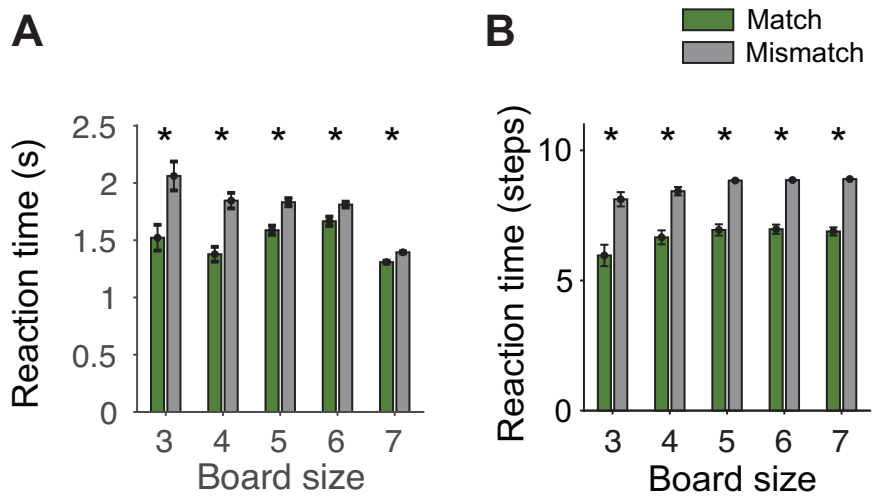


Figure 4.2: Reaction times for match (green) and mismatch (gray) trials. Figure shows metric for different board sizes for epilepsy patient participants (**A**) and Hebbian attractor (**B**) (Section 3.3). The reaction time of both the epilepsy patients and the Hebbian attractor of mismatch trials was longer than match trials. Error bars indicate s.e.m. (n=20 trials). Asterisks denote significant difference between match and mismatch trials (permutation test, 5,000 iterations, α=0.01).

Reaction times Fig. 4.2 showcases reaction times for match and mismatch trials across various board sizes for both epilepsy patient participants and our Hebbian attractor model. The reaction time for both the epilepsy patients and the model was longer for mismatch trials compared to match trials across all board sizes, indicating a significant difference ($p<0.001$, compare Fig. 4.2.A and B).

N-since-last-click In Fig. 4.3, we present the average NSLC values for the 2nd tile across different board sizes for both epilepsy patient participants and the Hebbian attractor model. These NSLC values increased with board size and were notably larger for mismatch trials compared to match trials, which aligns with the behavior observed in participants ($p<0.001$, compare Fig. 4.3.A and B).

N-since-pair Fig. 4.4 displays the average n-since-pair (NSP) values for the 2nd tile across various board sizes for both epilepsy patient participants and our Hebbian attractor model. Similar to NSLC, the NSP values increased with board size and were significantly larger in mismatch trials compared to match trials for all board sizes ($p<0.001$, compare Fig. 4.4.A and B). These findings demonstrate that our computational model replicates key aspects of human behavior in this context.

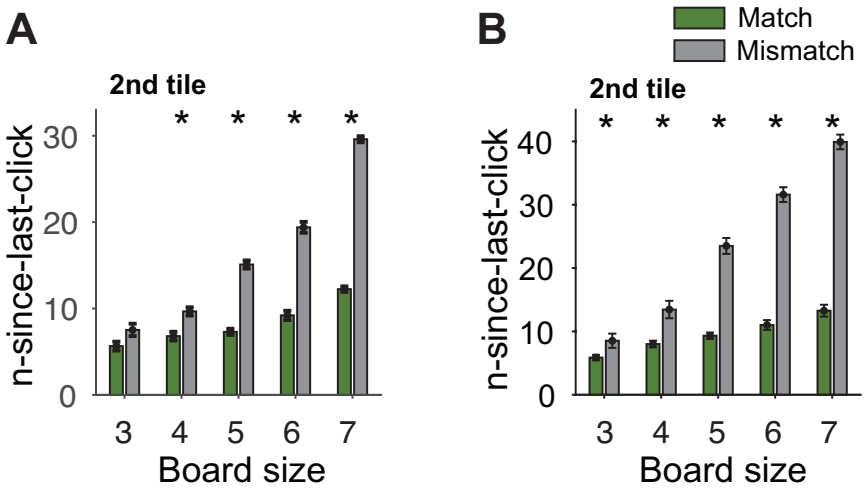


Figure 4.3: Average n-since-last-click for the 2nd tile for each board size. Figure shows metric for epilepsy patient participants (**A**) and Hebbian attractor (**B**) (Section 3.3). Error bars indicate s.e.m. (n=20 trials). Asterisks denote significant difference between match and mismatch trials (permutation test, 5,000 iterations, α=0.01).

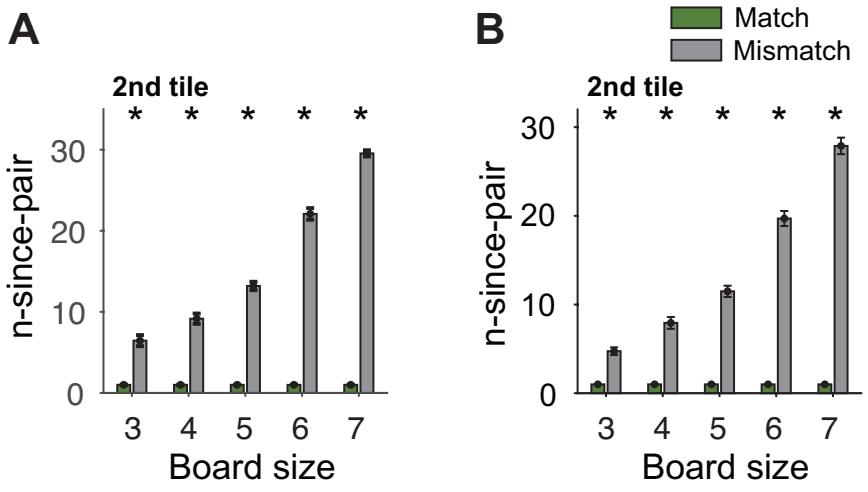


Figure 4.4: Average n-since-pair for the 2nd tile for each board size. Figure shows metric for epilepsy patient participants (**A**) and Hebbian attractor (**B**) (Section 3.3). Error bars indicate s.e.m. (n=20 trials). Asterisks denote significant difference between match and mismatch trials (permutation test, 5,000 iterations, α=0.01).

4.2.2 The model maps to the intracranial field potentials

To investigate the model's inner workings, we defined two metrics based on the unit activations.

Novelty/Familiarity signal To compare with the match related signals in Fig. 4.5.A, we computed an overall maximum energy (Section 3.4.6, Eq. (3.3)). This maximum energy was smaller for trials with NSLC=1 ($p<0.001$, Fig. 4.5.B), reflecting a strong correlate of memory for recently seen tiles (compare Fig. 4.5.A and B).

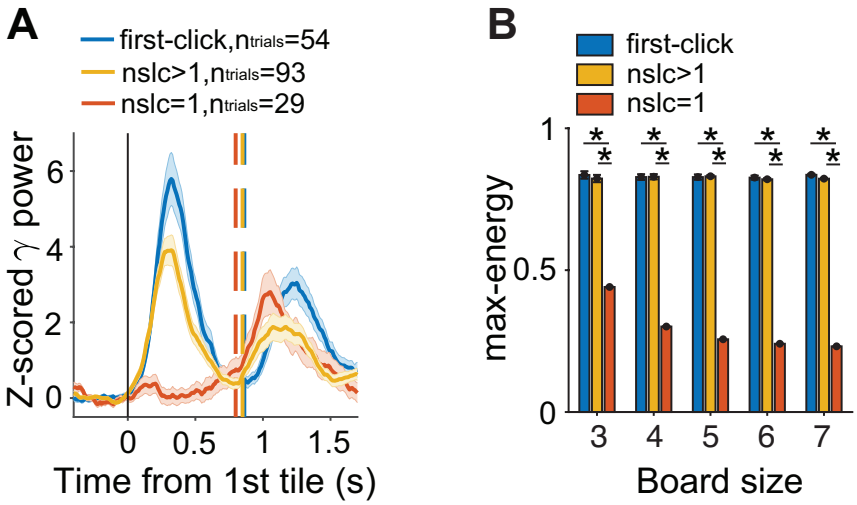


Figure 4.5: Comparison between novelty/familirity signal and model. A. Neural signal from the selected electrode in the left pars opercularis. The plot shows the Z-scored gamma band power aligned to the 1st tile onset (solid blue vertical line) for novel tiles (blue), unfamiliar tiles (n-since-last-click>1, yellow), and familiar tiles (n-since-last-click=1, red). The vertical dashed line indicates the mean reaction time. Multiple dashed lines indicate reaction time equalization (Section 3.4). The time axis extends from 400 ms before the click to 500 ms after the average reaction time. Shaded error bars indicate s.e.m. **B.** Max-energy for novel tiles (blue), unfamiliar tiles (n-since-last-click>1, yellow), and familiar tiles (n-since-last-click=1, red) (Section 3.4.6). Error bars indicate s.e.m. (n=20 trials). Asterisks denote significant difference between match and mismatch trials (permutation test, 5,000 iterations, α=0.01).

Confidence signal To compare with the match-related signals in Fig. 4.6.A, we defined a confidence metric by assessing the relative activation for the strongest unit with respect to the other units during the inference step (Section 3.4.6). The confidence metric was significantly larger for match trials compared to non-match trials ($p<0.01$, Fig. 4.6.B), which was qualitatively similar to the neural responses (compare Fig. 4.6.A and B).

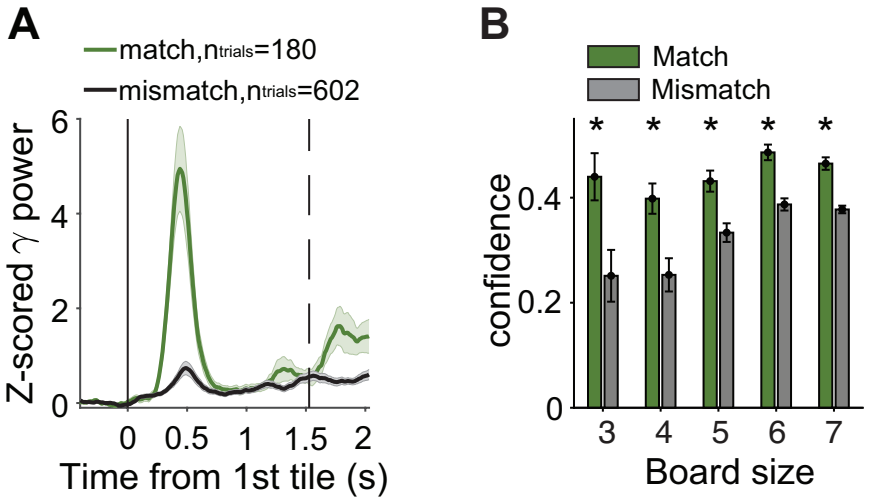


Figure 4.6: Comparison between confidence signal and model. A. Neural signal from the selected electrode in the right orbitofrontal cortex. The plot shows the Z-scored gamma band power aligned to the 1st tile onset (solid vertical line) for match trials (green) and mismatch trials (black). The vertical dashed line indicates the mean reaction time (Section 3.4). Shaded error bars indicate s.e.m. **B.** Model confidence for match (green) and mismatch (gray) trials for different board sizes (Section 3.4.6). Error bars indicate s.e.m. (n=20 trials). Asterisks denote significant difference between match and mismatch trials (permutation test, 5,000 iterations, α=0.01). The model confidence in match trials was larger than in mismatch trials.

Chapter 5

Discussion

5.1 Summary

We modeled behavioral and neurophysiological data collected from humans playing a natural task that involved working memory. The model we propose is a simple Hebbian attractor network, which is selected to match the precision of humans solving this task. From this simple model, we obtain accurate predictions on behavioral metrics beyond the precision, including memory strength and decay, and their effect on retrieval of past information. Moreover, from the same model's activation, we can extract signals that qualitatively map to iEEG recordings collected during the task.

5.1.1 Working memory in complex human behavior

In this study, we conducted an investigation into the neural dynamics during a memory task involving a classic card-matching game. Participants demonstrated competent task performance which was slightly below the level of a perfect memory model, as expected. Notably, they exhibited increased reaction times during mismatch trials and a discernible decay in memory traces over time since encoding.

This novel task strikes a balance between traditional memory studies involving sequentially presented stimuli and real-world behavioral scenarios [2, 8, 13, 20, 25, 113–119]. It introduces a complex yet realistic setting encompassing both associative and non-associative memory components. Despite this complexity, the task provides a high degree of experimental control over stimulus timing and parameters, which can be challenging to achieve in real-world memory research.

5.1.2 Intracranial recordings

To dissect the interplay of various intercorrelated variables inherent to complex and natural tasks, we employed a GLM. This model allowed us to quantitatively assess the influence of distinct predictors on neural responses, effectively capturing neurophysiological responses while considering predictor correlations. Additionally, extensive brain region sampling, involving data from over 1,000 electrodes across 20 participants, provided comprehensive coverage of neural responses during each task step.

The task initially involved recognizing whether a tile was novel or familiar, reflecting non-associative recognition memory. Strong neural responses signaling both novelty and familiarity were observed, particularly in areas like the lateral orbitofrontal cortex, the pars opercularis, and the medial temporal lobe. These responses were not content-specific, emphasizing the rarity of responses sharply tuned to specific sensory features in memory formation.

Following the first tile presentation, participants made internal predictions about whether they remembered the pair's location, influencing their subsequent click decisions. Neural responses notably reflected these predictions and the internal memory strength estimates or confidence levels. Even before the second tile revelation, significant neural differences between match and mismatch trials emerged. Transient and sustained responses were observed, potentially signifying sudden realizations, high confidence, and active retrieval processes. These findings emphasize the role of the hippocampus, medial temporal lobe, and other regions like the lateral orbitofrontal cortex in working memory.

5.1.3 A Hebbian attractor can predict human behavior

Building on previous research [43, 45, 82, 120], we provided a proof-of-principle demonstration showcasing the ability of a simple instantiation of such a model to qualitatively replicate human behavioral and neural responses in a complex memory task. Through a series of performance evaluators mirroring human behavior, our model not only replicated the pattern of increased clicks per tile with board size but also revealed prolonged reaction times during mismatch trials compared to match trials. Importantly, we introduced metrics based on unit activations to delve into the model's inner workings. These metrics, including an overall maximum energy reflecting memory for recently seen tiles and a confidence metric mirroring neural responses, provided further insights into the cognitive processes underpinning the task.

The implications of our work extend beyond the immediate findings, as this basic neural network architecture can be seamlessly integrated with visual neural networks, opening avenues to investigate the representation of visual

signals in working memory further. Moreover, the adaptability of our model allows for its extension to more intricate tasks, involving multi-way associations and dynamic environmental changes over time. The comprehensive analysis of human behavior also yielded valuable insights, highlighting the interplay between memory load, reaction times, and eye movements in a naturalistic setting.

These observations collectively represent initial steps towards advancing our understanding of the intricate interactions involved in the formation of natural memory events. By bridging the gap between computational modeling and real-world memory tasks, this thesis contributes to the broader field of memory research and sets the stage for further exploration into the complexities of memory encoding and retrieval mechanisms.

5.2 Future Work

The work presented in this thesis can be extended along different axes. For one, it would be interesting to explore how the same architecture could predict the first click selection from human behavior.

Another direction in which this work could be extended is the implementation of a visual/spatial backbone that would substitute the simple encoding of the image label and the tile position we analyzed in this work. For example, the number of units in the model could be scaled to match the size of a selected layer in a deep neural network, effectively allowing the model to memorize feature associations between positional and visual information. This line of work could be complemented by additional behavioral studies, that could include controlled similarity between images, or different arrangements of the tiles.

Moreover, the model could be tested on variations of the task we presented, including different grid configurations to study positional effects or tile-swapping to test the robustness of the model to noise. Additionally, this line of work could be extended by testing the model on different tasks, e.g. delayed match to sample or recall of past seen objects given partial information.

Lastly, it would be interesting to explore more behavioral metrics that could give interesting insights into the patients' and model's behavior. As an example, primacy and recency effects have not been studied in this work and could add an interesting contribution to finding differences or similarities between the patients and the model.

Appendix A

Algorithms

```
1  n ← number of units
2
3  t ← 0
4  M(t) ← rand(0,1)_{n×n}
5  h(t) ← [0]_n
6
7  function update_memory(x, steps)
8
9     while t < t+steps
10       h(t+1) ← σ(N(x + h(t)·M(t)))
11       M(t+1) ← λ·M(t) + η·h(t+1)·M(t)
12       t ← t+1
13    end while
14
15    return h(t)
16
17 end function
```

Algorithm A.1: Algorithm for updating the memory model as explained in Section 3.2.

```
p ← number of tiles
l ← number of labels

steps ← 10

function learn(memory, position, label)

    xp ← [-1]_p
    xl ← [-1]_l

    xp[position] ← 1
    xl[label] ← 1
    x ← [xp, xl]

    memory.update_memory(x, steps)

end function
```

Algorithm A.2: Algorithm for the learning regime explained in Section 3.2.2.

```
p ← number of tiles
l ← number of labels

steps ← 10

function infer(memory, available, label)

    xp ← [-1]_p
    xl ← [0]_l

    xp[available] ← 0
    xl[label] ← 1
    x ← [xp, xl]

    memory.update_memory(x, steps)

    i ← argmax(memory.p[available])
    return i

end function
```

Algorithm A.3: Algorithm for the inference regime explained in Section 3.2.2.

Appendix B

Frequencies of clicks per tile position

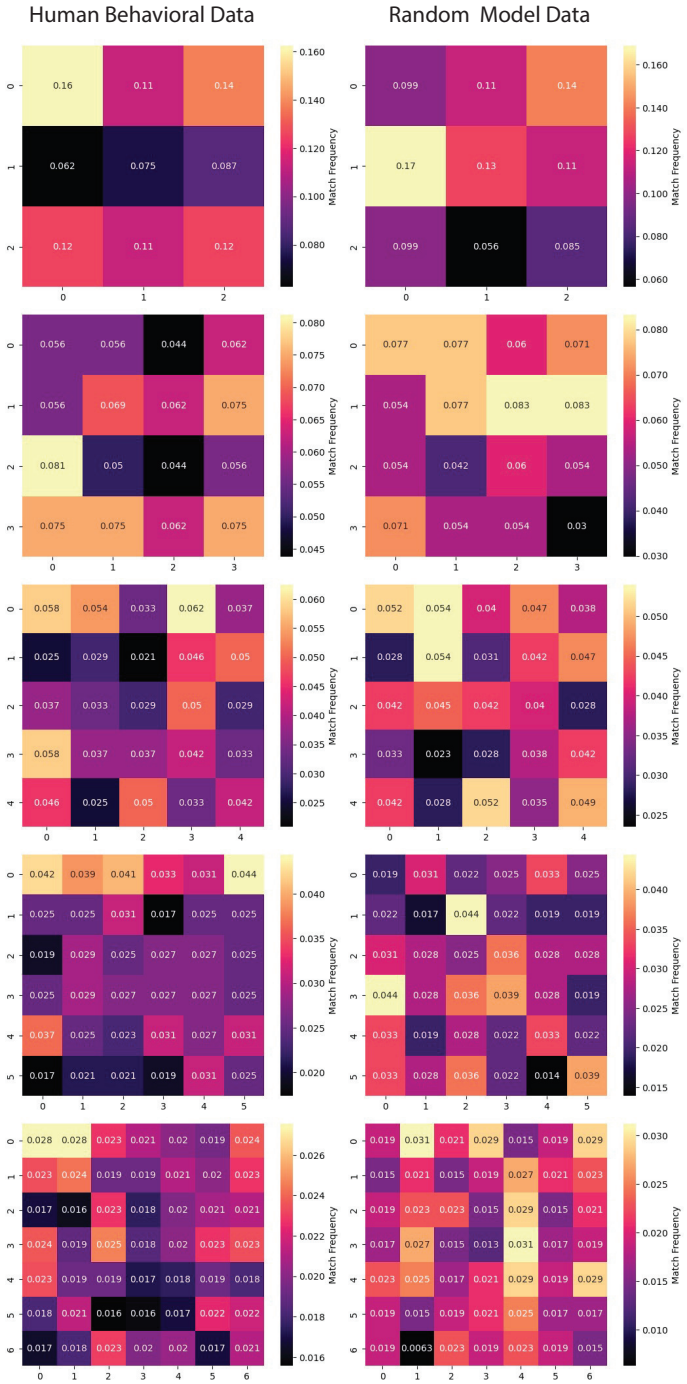


Figure B.1: Frequencies of clicks per tile position in match trials. Each row refers to one board size (3x3 on the top row to 7x7 on the bottom row). The left column shows the data from the human experiments. The right column shows the data from a random model.

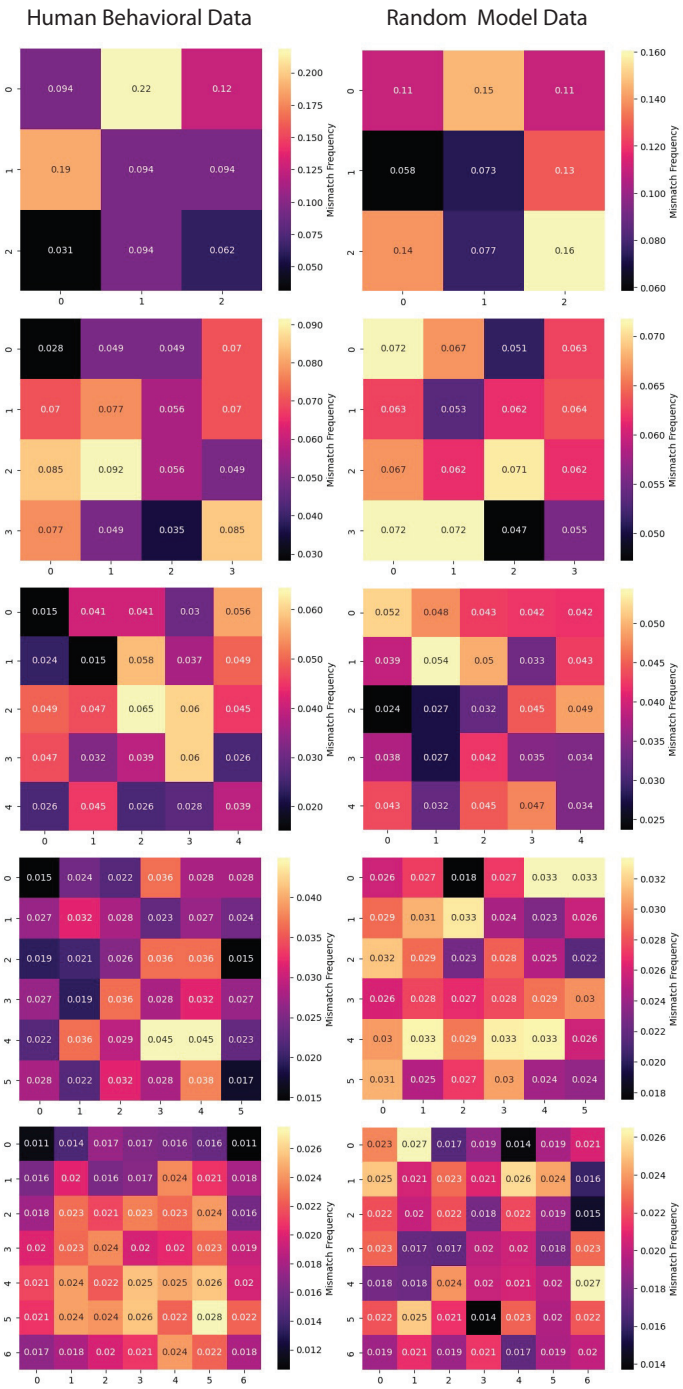


Figure B.2: Frequencies of clicks per tile position in mismatch trials. Each row refers to one board size (3x3 on the top row to 7x7 on the bottom row). The left column shows the data from the human experiments. The right column shows the data from a random model.

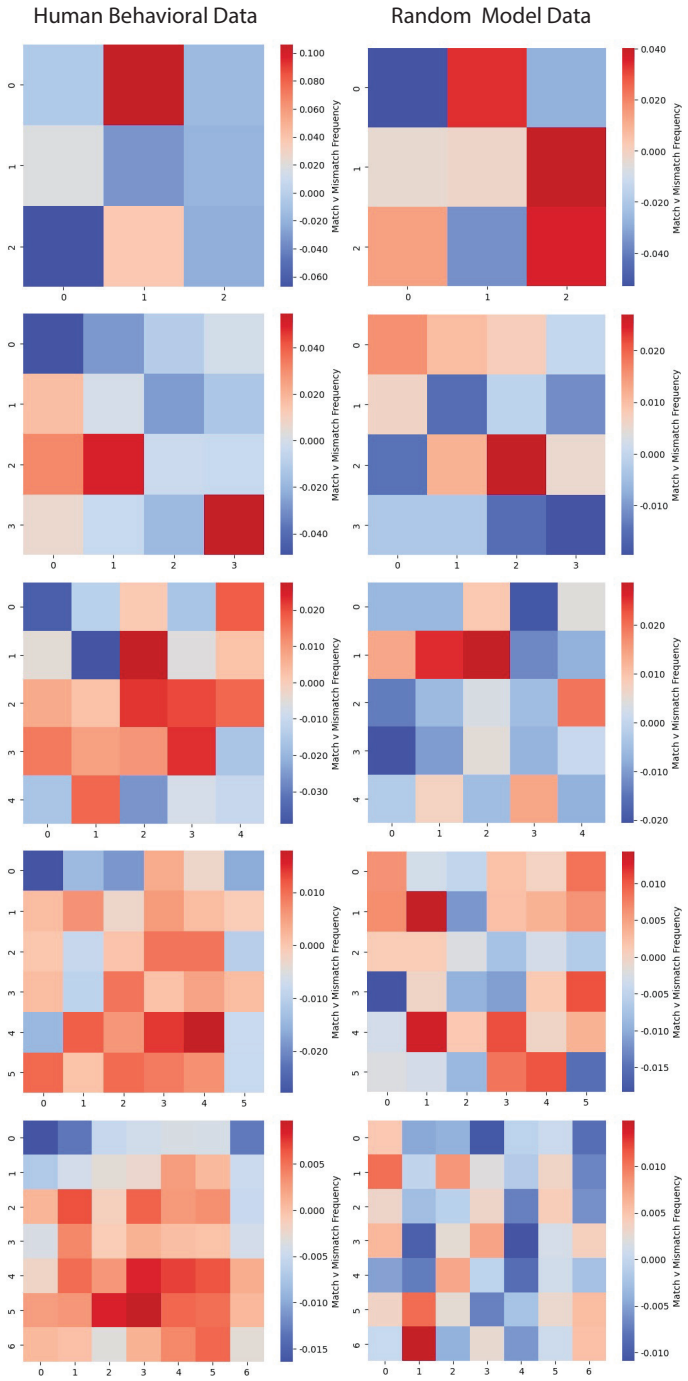


Figure B.3: Difference in frequencies of clicks per tile position between match vs mismatch trials. Each row refers to one board size (3x3 on the top row to 7x7 on the bottom row). The left column shows the data from the human experiments. The right column shows the data from a random model. Positive values indicate a higher frequency of mismatches, negative values indicate a higher frequency of matches.

Bibliography

[1] Yuchen Xiao, Paula Sánchez López, Ruijie Wu, Peng-Hu Wei, Yong-Zhi Shan, Daniel Weisholtz, Garth Rees Cosgrove, Joseph R Madsen, Scellig Stone, Guo-Guang Zhao, and Gabriel Kreiman. Integration of recognition, episodic, and associative memories during complex human behavior. *bioRxiv*, 2023. doi: 10.1101/2023.03.27.534384. URL https://www.biorxiv.org/content/early/2023/03/27/2023.03.27.534384.

[2] Endel Tulving and Neal Kroll. Novelty assessment in the brain and long-term memory encoding. *Psychonomic bulletin & review*, 2(3):387–390, 1995.

[3] Katherine D Duncan and Daphna Shohamy. Memory states influence value-based decisions. *Journal of Experimental Psychology: General*, 145 (11):1420, 2016.

[4] Daniela Montaldi, Tom J Spencer, Neil Roberts, and Andrew R Mayes. The neural system that mediates familiarity memory. *Hippocampus*, 16 (5):504–520, 2006.

[5] Vahid Mehrpour, Travis Meyer, Eero P Simoncelli, and Nicole C Rust. Pinpointing the neural signatures of single-exposure visual recognition memory. *Proceedings of the National Academy of Sciences*, 118(18): e2021660118, 2021.

[6] Jinsick Park, Hojong Lee, Taekyung Kim, Ga Young Park, Eun Mi Lee, Seunghee Baek, Jeonghun Ku, In Young Kim, Sun I Kim, Dong Pyo Jang, et al. Role of low-and high-frequency oscillations in the human hippocampus for encoding environmental novelty during a spatial navigation task. *Hippocampus*, 24(11):1341–1352, 2014.

[7] Michael A Yassa and Craig EL Stark. Multiple signals of recognition memory in the medial temporal lobe. *Hippocampus*, 18(9):945–954, 2008.

[8] Ueli Rutishauser, Ian B Ross, Adam N Mamelak, and Erin M Schuman. Human memory strength is predicted by theta-frequency phase-locking of single neurons. *Nature*, 464(7290):903–907, 2010.

[9] Sander M Daselaar, Mathias S Fleck, and R Cabeza. Triple dissociation in the medial temporal lobes: recollection, familiarity, and novelty. *Journal of neurophysiology*, 96(4):1902–1911, 2006.

[10] Itzhak Fried, Katherine A MacDonald, and Charles L Wilson. Single neuron activity in human hippocampus and amygdala during recognition of faces and objects. *Neuron*, 18(5):753–765, 1997.

[11] Ryan J Murray, Tobias Brosch, and David Sander. The functional profile of the human amygdala in affective processing: insights from intracranial recordings. *Cortex*, 60:10–33, 2014.

[12] Tino Zaehle, Eva M Bauch, Hermann Hinrichs, Friedhelm C Schmitt, Jürgen Voges, Hans-Jochen Heinze, and Nico Bunzeck. Nucleus accumbens activity dissociates different forms of salience: evidence from human intracranial recordings. *Journal of Neuroscience*, 33(20):8764–8771, 2013.

[13] Indre V Viskontas, Barbara J Knowlton, Peter N Steinmetz, and Itzhak Fried. Differences in mnemonic processing by neurons in the human hippocampus and parahippocampal regions. *Journal of cognitive neuroscience*, 18(10):1654–1662, 2006.

[14] Malcolm W Brown and John P Aggleton. Recognition memory: what are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience*, 2(1):51–61, 2001.

[15] Michal T Kucewicz, Jan Cimbalnik, Joseph Y Matsumoto, Benjamin H Brinkmann, Mark R Bower, Vincent Vasoli, Vlastimil Sulc, Fred Meyer, WR Marsh, SM Stead, et al. High frequency oscillations are associated with cognitive processing in human recognition memory. *Brain*, 137 (8):2231–2244, 2014.

[16] Jie Zheng, Andrea GP Schjetnan, Mar Yebra, Bernard A Gomes, Clayton P Mosher, Suneil K Kalia, Taufik A Valiante, Adam N Mamelak, Gabriel Kreiman, and Ueli Rutishauser. Neurons detect cognitive boundaries to structure episodic memories in humans. *Nature neuroscience*, 25(3):358–368, 2022.

[17] Ueli Rutishauser. Testing models of human declarative memory at the single-neuron level. *Trends in cognitive sciences*, 23(6):510–524, 2019.

[18] Elizabeth L Johnson and Robert T Knight. Intracranial recordings and human memory. *Current opinion in Neurobiology*, 31:18–25, 2015.

[19] Robert T Knight. Contribution of human hippocampal region to novelty detection. *Nature*, 383(6597):256–259, 1996.

[20] Per B Sederberg, Andreas Schulze-Bonhage, Joseph R Madsen, Edward B Bromfield, David C McCarthy, Armin Brandt, Michele S Tully, and Michael J Kahana. Hippocampal and neocortical gamma oscillations predict memory formation in humans. *Cerebral cortex*, 17(5): 1190–1196, 2007.

[21] Henry L Roediger III and Eylul Tekin. Recognition memory: Tulving's contributions and some new findings. *Neuropsychologia*, 139:107350, 2020.

[22] Matias J Ison, Rodrigo Quian Quiroga, and Itzhak Fried. Rapid encoding of new memories by individual neurons in the human brain. *Neuron*, 87(1):220–230, 2015.

[23] Timothy C Sheehan, Vishnu Sreekumar, Sara K Inati, and Kareem A Zaghloul. Signal complexity of human intracranial eeg tracks successful associative-memory formation across individuals. *Journal of Neuroscience*, 38(7):1744–1755, 2018.

[24] Sylvia Wirth, Marianna Yanike, Loren M Frank, Anne C Smith, Emery N Brown, and Wendy A Suzuki. Single neurons in the monkey hippocampus and learning of new associations. *Science*, 300(5625): 1578–1581, 2003.

[25] C Brock Kirwan and Craig EL Stark. Medial temporal lobe activation during encoding and retrieval of novel face-name pairs. *Hippocampus*, 14(7):919–930, 2004.

[26] Charan Ranganath, Michael X Cohen, Cathrine Dam, and Mark D'Esposito. Inferior temporal, prefrontal, and hippocampal contributions to visual working memory maintenance and associative memory retrieval. *Journal of Neuroscience*, 24(16):3917–3925, 2004.

[27] Kuniyoshi Sakai and Yasushi Miyashita. Neural organization for the long-term memory of paired associates. *Nature*, 354(6349):152–155, 1991.

[28] Yong-Di Zhou, Allen Ardestani, and Joaquín M Fuster. Distributed and associative working memory. *Cerebral Cortex*, 17(suppl_1):i77–i87, 2007.

[29] Ueli Rutishauser, Leila Reddy, Florian Mormann, and Johannes Sarnthein. The architecture of human memory: insights from human single-neuron recordings. *Journal of Neuroscience*, 41(5):883–890, 2021.

[30] Andrew Mayes, Daniela Montaldi, and Ellen Migo. Associative memory and the medial temporal lobes. *Trends in cognitive sciences*, 11(3): 126–135, 2007.

[31] Cyma Van Petten, Barbara J Luka, Susan R Rubin, and John P Ryan. Frontal brain activity predicts individual performance in an associative memory exclusion test. *Cerebral Cortex*, 12(11):1180–1192, 2002.

[32] Alan Baddeley. Working memory. *Current Biology*, 20(4):R136–R140, 2010. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2009.12.014. URL https://www.sciencedirect.com/science/article/pii/S0960982209021332.

[33] Omri Barak and Misha Tsodyks. Working models of working memory. *Current opinion in neurobiology*, 25:20–24, 2014.

[34] Albert Compte, Nicolas Brunel, Patricia S Goldman-Rakic, and Xiao-Jing Wang. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral cortex*, 10(9):910–923, 2000.

[35] Clayton E Curtis and Mark D'Esposito. Persistent activity in the prefrontal cortex during working memory. *Trends in cognitive sciences*, 7 (9):415–423, 2003.

[36] Renato Duarte, Alexander Seeholzer, Karl Zilles, and Abigail Morrison. Synaptic patterning and the timescales of cortical dynamics. *Current Opinion in Neurobiology*, 43:156–165, 2017.

[37] Daniel Durstewitz, Jeremy K Seamans, and Terrence J Sejnowski. Neurocomputational models of working memory. *Nature neuroscience*, 3 (11):1184–1191, 2000.

[38] Shintaro Funahashi, Charles J Bruce, and Patricia S Goldman-Rakic. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of neurophysiology*, 61(2):331–349, 1989.

[39] Joaquin M Fuster and Garrett E Alexander. Neuron activity related to short-term memory. *Science*, 173(3997):652–654, 1971.

[40] Katsuki Nakamura and KISOU Kubota. Mnemonic firing of neurons in the monkey temporal pole during a visual recognition memory task. *Journal of neurophysiology*, 74(1):162–178, 1995.

[41] Kei Watanabe and Shintaro Funahashi. Prefrontal delay-period activity reflects the decision process of a saccade direction during a free-choice odr task. *Cerebral Cortex*, 17(suppl_1):i88–i100, 2007.

[42] Rita Almeida, João Barbosa, and Albert Compte. Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. *Journal of neurophysiology*, 114(3):1806–1818, 2015.

[43] Anders Lansner, Petter Marklund, Sverker Sikström, and Lars-Göran Nilsson. Reactivation in working memory: an attractor network model of free recall. *PloS one*, 8(8):e73776, 2013.

[44] Julian Macoveanu, T Klingberg, and Jesper Tegnér. A biophysical model of multiple-item working memory: a computational and neuroimaging study. *Neuroscience*, 141(3):1611–1618, 2006.

[45] Sanjay Manohar, Nahid Zokaei, Sean Fallon, Tim Vogels, and Masud Husain. Neural mechanisms of attending to items in working memory. *Neuroscience & Biobehavioral Reviews*, 101, 03 2019. doi: 10.1016/j.neubiorev.2019.03.017.

[46] Chantal Roggeman, Torkel Klingberg, Heleen EM Feenstra, Albert Compte, and Rita Almeida. Trade-off between capacity and precision in visuospatial working memory. *Journal of Cognitive Neuroscience*, 26 (2):211–222, 2014.

[47] Alexander Seeholzer, Moritz Deger, and Wulfram Gerstner. Stability of working memory in continuous attractor networks under the control of short-term plasticity. *PLoS computational biology*, 15(4):e1006928, 2019.

[48] Ziqiang Wei, Xiao-Jing Wang, and Da-Hui Wang. From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. *Journal of Neuroscience*, 32(33): 11228–11240, 2012.

[49] Florian Fiebig and Anders Lansner. A spiking working memory model based on hebbian short-term potentiation. *Journal of Neuroscience*, 37 (1):83–96, 2017.

[50] Yuanyuan Mi, Mikhail Katkov, and Misha Tsodyks. Synaptic correlates of working memory capacity. *Neuron*, 93(2):323–330, 2017.

[51] Gianluigi Mongillo, Omri Barak, and Misha Tsodyks. Synaptic theory of working memory. *Science*, 319(5869):1543–1546, 2008.

[52] Sandro Romani and Misha Tsodyks. Short-term plasticity based network model of place cells dynamics. *Hippocampus*, 25(1):94–105, 2015.

[53] Lawrence Christopher York and Mark CW van Rossum. Recurrent networks with short term synaptic depression. *Journal of computational neuroscience*, 27(3):607–620, 2009.

[54] William James. The principles of psychology volume ii by william james (1890). 1890.

[55] Nelson Cowan. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological bulletin*, 104(2):163, 1988.

[56] Noelle Wood and Nelson Cowan. The cocktail party phenomenon revisited: how frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1):255, 1995.

[57] Nelson Cowan. An embedded-processes model of working memory. *Models of working memory: Mechanisms of active maintenance and executive control*, 20(506):1013–1019, 1999.

[58] Nelson Cowan. *Working memory capacity.* Essays in cognitive psychology. Psychology Press, New York, NY, US, 2005. ISBN 1-84169-097-X (Hardcover); 978-1-84169-097-1 (Hardcover). doi: 10.4324/9780203342398. URL https://doi.org/10.4324/9780203342398.

[59] Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1):87–114, 2001.

[60] R.C. Atkinson and R.M. Shiffrin. Human memory: A proposed system and its control processes11this research was supported by the national aeronautics and space administration, grant no. ngr-05-020-036. the authors are indebted to w. k. estes and g. h. bower who provided many valuable suggestions and comments at various stages of the work. special credit is due j. w. brelsford who was instrumental in carrying out the research discussed in section iv and whose overall contributions are too numerous to report in detail. we should also like to thank those co-workers who carried out a number of the experiments discussed in the latter half of the paper; rather than list them here, each will be acknowledged at the

appropriate place. volume 2 of *Psychology of Learning and Motivation*, pages 89–195. Academic Press, 1968. doi: https://doi.org/10.1016/S0079-7421(08)60422-3. URL https://www.sciencedirect.com/science/article/pii/S0079742108604223.

[61] Nelson Cowan. What are the differences between long-term, short-term, and working memory? *Progress in brain research*, 169:323–338, 2008.

[62] Alan D. Baddeley and Graham Hitch. Working memory. volume 8 of *Psychology of Learning and Motivation*, pages 47–89. Academic Press, 1974. doi: https://doi.org/10.1016/S0079-7421(08)60452-1. URL https://www.sciencedirect.com/science/article/pii/S0079742108604521.

[63] Mark D'Esposito and Bradley Postle. The cognitive neuroscience of working memory. *Annual review of psychology*, 66, 09 2014. doi: 10.1146/annurev-psych-010814-015031.

[64] T. A. R. *The American Journal of Psychology*, 75(1):161–163, 1962. ISSN 00029556. URL http://www.jstor.org/stable/1419559.

[65] KH Pribham, A Ahumada, J Hartog, and L Roos. The frontal cortex and behavior. 1964.

[66] Klaus Wimmer, Duane Q Nykamp, Christos Constantinidis, and Albert Compte. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature neuroscience*, 17(3):431–439, 2014.

[67] Javier Quintana, Joaquin M Fuster, and Javier Yajeya. Effects of cooling parietal cortex on prefrontal units in delay tasks. *Brain research*, 503(1):100–110, 1989.

[68] Nicolas Y Masse, Guangyu R Yang, H Francis Song, Xiao-Jing Wang, and David J Freedman. Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature neuroscience*, 22(7):1159–1167, 2019.

[69] Kei Watanabe and Shintaro Funahashi. Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. *Nature neuroscience*, 17(4):601–611, 2014.

[70] Mark G Stokes. 'activity-silent'working memory in prefrontal cortex: a dynamic coding framework. *Trends in cognitive sciences*, 19(7):394–405, 2015.

[71] Mikael Lundqvist, Jonas Rose, Pawel Herman, Scott L Brincat, Timothy J Buschman, and Earl K Miller. Gamma and beta bursts underlie working memory. *Neuron*, 90(1):152–164, 2016.

[72] Jan Kamiński and Ueli Rutishauser. Between persistently active and activity-silent frameworks: novel vistas on the cellular basis of working memory. *Annals of the New York Academy of Sciences*, 1464(1):64–75, 2020.

[73] Donald O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York, 1949.

[74] Tim VP Bliss and Graham L Collingridge. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, 361(6407): 31–39, 1993.

[75] Robert C Malenka. Postsynaptic factors control the duration of synaptic enhancement in area ca1 of the hippocampus. *Neuron*, 6(1):53–60, 1991.

[76] Martha A Erickson, Lauren A Maramara, and John Lisman. A single brief burst induces glur1-dependent associative short-term potentiation: a potential mechanism for short-term memory. *Journal of cognitive neuroscience*, 22(11):2530–2540, 2010.

[77] Pojeong Park, Arturas Volianskis, Thomas M Sanderson, Zuner A Bortolotto, David E Jane, Min Zhuo, Bong-Kiun Kaang, and Graham L Collingridge. Nmda receptor-dependent long-term potentiation comprises a family of temporally overlapping forms of synaptic plasticity that are induced by different patterns of stimulation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1633): 20130131, 2014.

[78] Arturas Volianskis, Grace France, Morten S Jensen, Zuner A Bortolotto, David E Jane, and Graham L Collingridge. Long-term potentiation and the role of n-methyl-d-aspartate receptors. *Brain research*, 1621:5–16, 2015.

[79] David W Tank and John J Hopfield. Collective computation in neuronlike circuits. *Scientific American*, 257(6):104–115, 1987.

[80] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[81] Mrazvan22. Plagiarism — Wikipedia, the free encyclopedia, 2013. URL https://upload.wikimedia.org/wikipedia/commons/4/49/Energy_landscape.png. [Online; accessed 9-Sep-2023].

[82] Mikail Khona and Ila R Fiete. Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, 23(12):744–766, 2022.

[83] Xiao-Jing Wang. Attractor network models. In *Encyclopedia of neuroscience*, pages 667–679. Elsevier Ltd, 2009.

[84] Jeffrey P Mullin, Michael Shriver, Soha Alomar, Imad Najm, Juan Bulacio, Patrick Chauvel, and Jorge Gonzalez-Martinez. Is seeg safe? a systematic review and meta-analysis of stereo-electroencephalography–related complications. *Epilepsia*, 57(3):386–401, 2016.

[85] Josef Parvizi and Sabine Kastner. Promises and limitations of human intracranial electroencephalography. *Nature neuroscience*, 21(4): 474–483, 2018.

[86] Ole Jensen, Jochen Kaiser, and Jean-Philippe Lachaux. Human gamma-frequency oscillations associated with attention and memory. *Trends in neurosciences*, 30(7):317–324, 2007.

[87] Peter J Uhlhaas, Gordon Pipa, Sergio Neuenschwander, Michael Wibral, and Wolf Singer. A new look at gamma? high-(¿ 60 hz) γ-band activity in cortical networks: function, mechanisms and impairment. *Progress in biophysics and molecular biology*, 105(1-2):14–28, 2011.

[88] György Buzsáki, Costas A Anastassiou, and Christof Koch. The origin of extracellular fields and currents—eeg, ecog, lfp and spikes. *Nature reviews neuroscience*, 13(6):407–420, 2012.

[89] Jessica A Cardin, Marie Carlén, Konstantinos Meletis, Ulf Knoblich, Feng Zhang, Karl Deisseroth, Li-Huei Tsai, and Christopher I Moore. Driving fast-spiking cells induces gamma rhythm and controls sensory responses. *Nature*, 459(7247):663–667, 2009.

[90] Nathan E Crone, Anna Korzeniewska, and Piotr J Franaszczuk. Cortical gamma responses: searching high and low. *International Journal of Psychophysiology*, 79(1):9–15, 2011.

[91] Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.

[92] Jean-Philippe Lachaux, Nikolai Axmacher, Florian Mormann, Eric Halgren, and Nathan E Crone. High-frequency neural activity and human cognition: past, present and possible future of intracranial eeg research. *Progress in neurobiology*, 98(3):279–301, 2012.

[93] Mikael Lundqvist, Pawel Herman, Melissa R Warden, Scott L Brincat, and Earl K Miller. Gamma and beta bursts during working memory readout suggest roles in its volitional control. *Nature communications*, 9(1):394, 2018.

[94] Earl K Miller, Mikael Lundqvist, and André M Bastos. Working memory 2.0. *Neuron*, 100(2):463–475, 2018.

[95] Saskia Haegens, Verónica Nácher, Rogelio Luna, Ranulfo Romo, and Ole Jensen. α-oscillations in the monkey sensorimotor network influence discrimination performance by rhythmical inhibition of neuronal spiking. *Proceedings of the National Academy of Sciences*, 108(48):19377–19382, 2011.

[96] Eelke Spaak, Mathilde Bonnefond, Alexander Maier, David A Leopold, and Ole Jensen. Layer-specific entrainment of gamma-band neural activity by the alpha rhythm in monkey visual cortex. *Current biology*, 22(24):2313–2318, 2012.

[97] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[98] David H. Brainard. The psychophysics toolbox. *Spatial Vision*, 10(4):433 – 436, 1997. doi: https://doi.org/10.1163/156856897X00357. URL https://brill.com/view/journals/sv/10/4/article-p433_15.xml.

[99] Denis G. Pelli. The videotoolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, 10(4):437 – 442, 1997. doi: https://doi.org/10.1163/156856897X00366. URL https://brill.com/view/journals/sv/10/4/article-p437_16.xml.

[100] Jimmy Ba, Geoffrey Hinton, Volodymyr Mnih, Joel Z. Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past, 2016.

[101] Itzhak Fried, Ueli Rutishauser, Moran Cerf, and Gabriel Kreiman. *Single Neuron Studies of the Human Brain: Probing Cognition*. The MIT Press, 07 2014. ISBN 9780262027205. doi: 10.7551/mitpress/9780262027205.001.0001. URL https://doi.org/10.7551/mitpress/9780262027205.001.0001.

[102] Rahul Desikan, Florent Ségonne, Bruce Fischl, Brian Quinn, Bradford Dickerson, Deborah Blacker, Randy Buckner, Anders Dale, Ralph Maguire, Bradley Hyman, Marilyn Albert, and Ronald Killiany. An

automated labeling system for subdiving the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, 31:968–80, 08 2006. doi: 10.1016/j.neuroimage.2006.01.021.

[103] David M. Groppe, Stephan Bickel, Andrew R. Dykstra, Xiuyuan Wang, Pierre Mégevand, Manuel R. Mercier, Fred A. Lado, Ashesh D. Mehta, and Christopher J. Honey. ielvis: An open source matlab toolbox for localizing and visualizing human intracranial electrode data. *Journal of Neuroscience Methods*, 281:40–48, 2017. ISSN 0165-0270. doi: https://doi.org/10.1016/j.jneumeth.2017.01.022. URL https://www.sciencedirect.com/science/article/pii/S0165027017300365.

[104] Anders M. Dale, Bruce Fischl, and Martin I. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999. ISSN 1053-8119. doi: https://doi.org/10.1006/nimg.1998.0395. URL https://www.sciencedirect.com/science/article/pii/S1053811998903950.

[105] Alark Joshi, Dustin Scheinost, Hirohito Okuda, Dominique Belhachemi, Isabella Murphy, Lawrence Staib, and Xenophon Papademetris. Unified framework for development, deployment and robust testing of neuroimaging algorithms. *Neuroinformatics*, 9:69–84, 03 2011. doi: 10.1007/s12021-010-9092-8.

[106] Jianxiao Wu, Gia Ngo, Douglas Greve, Li Jingwei, Tong He, Bruce Fischl, Simon Eickhoff, and B.T. Thomas Yeo. Accurate nonlinear mapping between mni volumetric and freesurfer surface coordinate systems. *Human Brain Mapping*, 39, 05 2018. doi: 10.1002/hbm.24213.

[107] Jiarui Wang, Annabelle Tao, William S. Anderson, Joseph R. Madsen, and Gabriel Kreiman. Mesoscopic physiological interactions in the human brain reveal small-world properties. *Cell Reports*, 36(8):109585, 2021. ISSN 2211-1247. doi: https://doi.org/10.1016/j.celrep.2021.109585. URL https://www.sciencedirect.com/science/article/pii/S2211124721010196.

[108] Arjun K. Bansal, Jedediah M. Singer, William S. Anderson, Alexandra Golby, Joseph R. Madsen, and Gabriel Kreiman. Temporal stability of visually selective responses in intracranial field potentials recorded from human occipital and temporal lobes. *Journal of Neurophysiology*, 108(11):3073–3086, 2012. doi: 10.1152/jn.00458.2012. URL https://doi.org/10.1152/jn.00458.2012. PMID: 22956795.

[109] Partha Mitra and Hemant Bokil. *Observed Brain Dynamics*. Oxford University Press, 12 2007. ISBN 9780195178081. doi: 10.1093/acprof:

oso/9780195178081.001.0001. URL https://doi.org/10.1093/acprof:oso/9780195178081.001.0001.

[110] Yuchen Xiao, Chien-Chen Chou, Garth Cosgrove, Nathan Crone, Scellig Stone, Joseph Madsen, Ian Reucroft, Yen-Cheng Shih, Daniel Weisholtz, Hsiang-Yu Yu, William Anderson, and Gabriel Kreiman. Cross-task specificity and within-task invariance of cognitive control processes. *Cell Reports*, 42:111919, 01 2023. doi: 10.1016/j.celrep.2022.111919.

[111] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384, 1972.

[112] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

[113] Reza Habib, Anthony R McIntosh, Mark A Wheeler, and Endel Tulving. Memory encoding and hippocampally-based novelty/familiarity discrimination networks. *Neuropsychologia*, 41(3):271–279, 2003.

[114] Yang Jiang, James V Haxby, Alex Martin, Leslie G Ungerleider, and Raja Parasuraman. Complementary neural mechanisms for tracking items in human working memory. *Science*, 287(5453):643–646, 2000.

[115] Tiffany E Chow and Jesse Rissman. Neurocognitive mechanisms of real-world autobiographical memory retrieval: insights from studies using wearable camera technology. *Annals of the New York Academy of Sciences*, 1396(1):202–221, 2017.

[116] Melissa C Duff, Tracey Wszalek, Daniel Tranel, and Neal J Cohen. Successful life outcome and management of real-world memory demands despite profound anterograde amnesia. *Journal of clinical and experimental neuropsychology*, 30(8):931–945, 2008.

[117] Dylan M Nielson, Troy A Smith, Vishnu Sreekumar, Simon Dennis, and Per B Sederberg. Human hippocampus represents space and time during retrieval of real-world memories. *Proceedings of the National Academy of Sciences*, 112(35):11078–11083, 2015.

[118] Pranav Misra, Alyssa Marconi, Matthew Peterson, and Gabriel Kreiman. Minimal memory for details in real life events. *Scientific reports*, 8(1):16701, 2018.

[119] Hanlin Tang, Jed Singer, Matias J Ison, Gnel Pivazyan, Melissa Romaine, Rosa Frias, Elizabeth Meller, Adrianna Boulin, James Carroll, Victoria Perron, et al. Predicting episodic memory formation for movie events. *Scientific reports*, 6(1):30175, 2016.

[120] Davide Spalla, Isabel Maria Cornacchia, and Alessandro Treves. Continuous attractors for dynamic memories. *Elife*, 10:e69499, 2021.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Hebbian attractor to model working memory in complex human behavior

Authored by (in block letters):

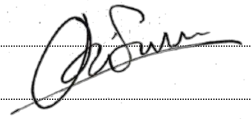
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Ravi	Srinivasan

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Milan, 26/09/2023	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.