Methods to Study Human Memory

Randolph F. Helfrich, M.D., Ph.D.[1], Robert T. Knight, M.D.[2,3], Mark D'Esposito, M.D.[2,3]

[1] *Hertie Institute for Clinical Brain Research, University of Tübingen, 72076 Tübingen, Germany*

[2] *Helen Wills Neuroscience Institute, UC Berkeley, Barker Hall, Berkeley, CA 94720, USA*

[3] *Dept. of Psychology, UC Berkeley, 2121 Berkeley Way, Berkeley, CA 94720, USA*

Contact Information

Randolph F. Helfrich, MD, PhD, Hertie Institute for Clinical Brain Research, University of Tübingen, 72076 Tübingen, Germany, phone: +49-7071-29-61898. email: randolph.helfrich@gmail.com

Robert T. Knight, MD, 130 Barker Hall, Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720-3190, USA, phone: +1-510-6430-9744. email: rtknight@berkeley.edu

Mark D'Esposito, MD, 132 Barker Hall, Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720-3190, USA, phone: +1-510-642-2839. email: despo@berkeley.edu

Abstract

A wide range of methods have advanced our understanding of the neural mechanisms underlying human memory function. For decades, the lesion approach served as the gold standard in localizing function and establishing causal relationships between anatomy and behavior. In the past 30 years, a wealth of evidence from neuroimaging (PET and functional MRI) and neurophysiological studies (MEG, scalp EEG, intracranial EEG and single unit recordings) has provided more detailed insights into the functional mechanisms of large-scale neuronal networks that enable memory formation. In addition, methodological advances in our ability to alter brain activity through electrical or magnetic stimulation has offered new insights into the role of such activity in causally modulating memory encoding, consolidation and retrieval. Here we review each of these methodological approaches and their strengths and weaknesses in addressing theoretical issues in memory research.

Keywords

Multimodal Imaging, Clinical Neuroscience, Neuropsychological Lesion Approach, Intracranial Electrophysiology, Activity-silent population coding, Static vs. dynamic coding, Sustained activity vs. transient bursts, Brain connectivity, Large-scale networks, Memory decoding

1. Introduction

Memory constitutes one of the most powerful cognitive faculties of the human brain. Memory formation and recall constitute complex processes that have experimentally and theoretically been divided into several core concepts and principles (Chapter 1/2), including encoding, (re-)consolidation and retrieval. Memory is typically grouped into declarative or procedural memory systems with further divisions made according to the assumed function including working memory and short-term and long-term memory. The theoretical combinations of these concepts already indicates that memory is a high-dimensional process with numerous facets. In this chapter, we review cognitive neuroscience approaches to study memory. We review several approaches, discuss their advantages and (dis-) advantages and highlight what a specific method contributes to our understanding in terms of 'where', 'when' and 'how' memory processing occurs in the human brain. Despite a surge of methods that range from single unit to whole-brain recordings at temporal resolution that range from μs to hours, no single method is able to fully address how all memory systems operate. Therefore, we also discuss evidence from multimodal investigations and review several novel analytical tools that integrate insights from multimodal imaging into the function of human memory. Note that we do not provide an exhaustive list of studies that have utilized a given method, but highlight several seminal findings that demonstrate how a specific method can be used to dissect the 'where', 'when' and 'how' of memory processes. We do not distinguish different memory concepts, such as declarative, procedural, episodic or working memory, but take a methods-centric perspective to describe how different aspects of memory can be studied using tools from cognitive neuroscience.

2. A convergent approach: Cognitive neuroscience tools to study human memory

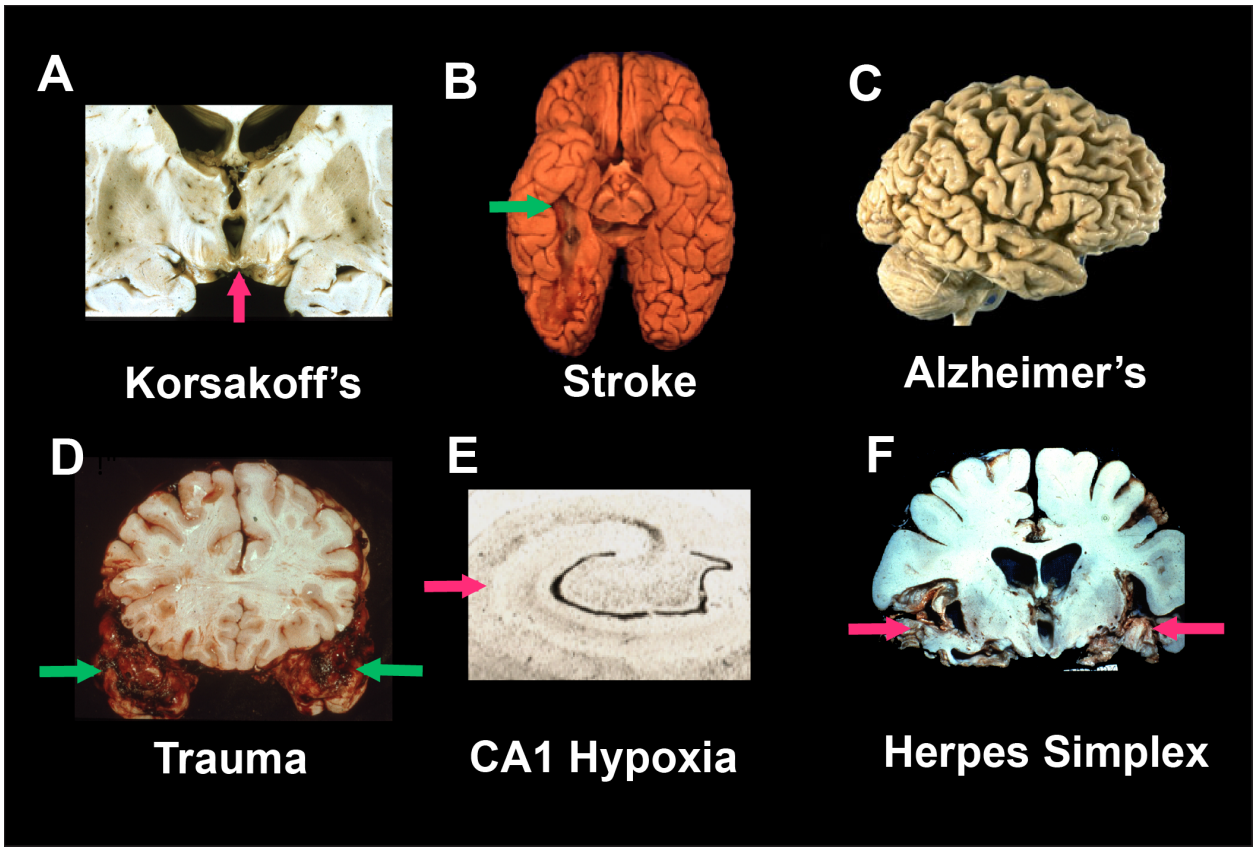
The study of human memory is inextricably associated with the neuropsychological lesion approach (Rorden & Karnath, 2004). Over the last few centuries, the systematic assessment of cognitive deficits in patients with circumscribed brain lesions has provided important insights into the functional organization of the human brain (Scoville & Milner, 1957; Szczepanski & Knight, 2014). Prior to modern day neuroimaging, this approach provided the only opportunity to establish a direct causal link between anatomy and cognition.

The history of human memory research might have taken a different turn if neurosurgeon William Scoville had not removed both hippocampi in one patient to treat his seizure disorder. While surgery reduced the number of seizures, it also left the patient permanently unable to form new memories. Over the next few decades, neuropsychologists Brenda Milner and Suzanne Corkin studied the case of Henry Molaison, or patient H. M., in great detail and their findings provided fundamental insights into how human memory systems are organized (Corkin, 2002; Scoville & Milner, 1957; Squire, 2009). For instance, they demonstrated that H. M. suffered from anterograde as well as retrograde amnesia impacting his episodic memory as well as new semantic learning, but his procedural memory and working memory systems remained largely intact.

However, there are numerous brain lesions outside the medial temporal lobe that can cause memory disorders (**Figure 1**). The detailed study of different patient cohorts has shown that memory capacity is distributed across large-scale cortical networks, with different nodes supporting distinct functions. For example, memory loss can also be caused by neocortical lesions (e.g. trauma, ischemic strokes), circumscribed subcortical lesions (e.g. degeneration of

mammillary bodies caused by nutritional thiamine deficiency resulting in Korsakoff syndrome), or neurodegenerative diseases that impact large-scale brain networks such as Alzheimer's (see Chapters 9.2 "Alzheimer's Disease Pathology and Cognition: Normal Aging to Clinical Dementia"). Hence, the etiology of memory disorders is diverse, and over the decades careful clinical observations provided the only means to develop a neurocognitive model of memory.

Figure 1



Memory Disorders. (A) Shrunken mammillary bodies (red arrow) in Korsakoff's amnesia due to thiamine deficiency. (B) Hippocampal infarction due to occlusion of the posterior cerebral artery (green arrow). (C) Global atrophy in Alzheimer's disease. Disease typically presents with anterograde amnesia due to initial pathology in the entorhinal cortex. (D) Bilateral hemorrhagic damage to both temporal lobes in a fatal traumatic brain injury (green arrows). Survivors have residual memory deficits. (E) Glutamate mediated loss of CA1 neurons due to hypoxia (red arrow). (F) Destruction of both temporal lobes in a survivor of Herpes Simplex encephalitis (red arrows).

There are numerous methods now available to study human memory, which differ in several important ways. While some methods are more suitable to localize function to a given anatomical structure, others offer the advantage of tracing the precise temporal evolution of neural processing that underlies memory formation.

Here, we group the most important methods into those that provide correlative or causal evidence. For instance, with the rise of modern day whole-brain neuroimaging by means of functional magnetic resonance imaging (fMRI) and positron-emission tomography (PET) in the 1990s, the lesion approach became less important as a localizer tool, but is still being used to establish causality as reviewed below.

2.1 Correlation vs. Causality

Correlative methods:

Functional MRI (fMRI): FMRI was introduced in the early 1990s. The fMRI BOLD (blood oxygenation level dependent) response can be modeled using a hemodynamic response function and typically needs 4-6s to reach peak magnitude after a brief experimental manipulation, i.e. a sensory stimulus or movement execution and takes several seconds to decay back to baseline (D'Esposito et al., 2009). Despite the sluggish response of the BOLD signal, even brief experimental events for as short as ~30ms can be detected with fMRI and different events that are spaced as closely as 500ms can be disentangled (Kim et al., 1997; Savoy, 1996; Zarahn et al., 1997). The latter requires the use of randomized inter-stimulus intervals as well as task designs that feature multiple conditions; otherwise discrete events need to be separated by at least 4s in a

fixed inter-stimulus design (Burock et al., 1998). The BOLD signal typically consists of two components: First, a brief 'initial dip' and a second large and more sustained BOLD increase, which is typically utilized to detect correlations with a behavioral task. It has previously been observed that the initial dip of the BOLD response likely indexes the actual site of neural activity more closely than the later positive portion of the BOLD response (Kim & Duong, 2002).

fMRI allows whole-brain imaging with high spatial resolution. Owing to its ability to detect whole-brain activity patterns, fMRI has enjoyed enormous popularity as a tool to study human memory. fMRI enabled testing and directly contrasting different cognitive models of memory in humans (e.g. activations of long-term memory representations, recruitment of sensory or motor areas, working memory capacity limits; D'Esposito & Postle, 2015). In contrast to the lesion approach, fMRI provided a more specific localization (e.g. dissecting hierarchical representations in the prefrontal cortex; Koechlin et al. 2003; Koechlin & Summerfield, 2007; Badre & D'Esposito, 2007) and mapping of e.g. subregions of the human prefrontal cortex (e.g. of the prefrontal cortex; Buckner & Petersen, 2006) or the hippocampus (Carr et al., 2010; Wisse et al., 2012). Note that the typical voxel size is approximately 3x3x3 mm on a 1.5 or 3T MRI scanner. However, at the expense of whole-brain coverage, sub-millimeter resolution can be obtained to image a single region in great detail. Using fMRI, researchers can account for structural and functional heterogeneity within a region and directly compare differences and/or commonalities of brain activations (Berman et al., 2006). Moreover, using fMRI inter-individual differences, influences of top-down cognitive control (Gazzaley et al., 2005) and compensatory activations in inter-connected networks can be studied (Rajah & D'Esposito, 2005).

Studies on 7T MRI scanners will further improve this spatial resolution and increase the extent of brain coverage (Shah et al., 2018; Thomas et al., 2008).. Furthermore, fMRI can be integrated

with other methods such as brain stimulation, EEG or pharmacological interventions (Mulert & Lemieux, 2009). However, fMRI also several limitations, including that it does not enable establishing causality, has low temporal resolution and is prone to artifacts from head movement, air sinuses and vascular pathology. Furthermore, fMRI only measures blood flow and not neural activity. An important caveat is that the neurovascular coupling (association of the BOLD signal and neuronal firing) is not fully understood (Logothetis, 2008).

Magneto- and Electroencephalography (M/EEG): M/EEG are non-invasive methods that either monitor the voltage fluctuations of the brains' electric field (EEG) or the corresponding, orthogonal magnetic field that is being produced by intracellular electric currents (MEG) (Baillet, 2017; Biasiucci et al., 2019; Cohen, 2017; Pesaran et al., 2018). Both methods have a high temporal resolution (milliseconds; typically sampled at >= 1000 Hz). The spatial resolution depends on the number of sensors being placed on or around the skull. Scalp EEG is typically recorded from as few as 19 sensors for clinical purposes, while research systems typically range between 64 and 256 channels offering a comparable number of sensors as state-of-the-art MEG (~300 sensors). The challenge of M/EEG is to infer where in the brain these electrical patterns emerge. The skull and skin are not ideal conductors and hence, contribute to signal attenuation in higher frequencies (> 30 Hz) and spatial spread of the signals since current follows the path of least resistance.

Although applying source modeling methods to high-density EEG and MEG data can suggest potential source generators (Pascual-Marqui et al., 1994; Pesaran et al., 2018; Van Veen et al., 1997). However, these methods attempt to solve an inverse problem that does not possess a

unique solution; that is, any given pattern of scalp-measured EEG/MEG signal could result from several underlying source configurations.

Over the last two decades, scalp EEG experienced a renaissance as new analytical tools became available (Biasiucci et al., 2019). Although the discovery of neuronal oscillations (Berger, 1929) emerged from early scalp EEG studies, investigation of these oscillations, and their role in memory function, received virtually no attention until the 1990s (Klimesch 1999; Tallon-Baudry and Bertrand, 1999; Tesche and Karhu, 2000; Caplan et al., 2001; Kahana et al. 2001; Kahana et al., 2006). Instead, EEG analyses focused on the study of stimulus and response-locked neural activity, referred to as event-related potentials (ERPs, requiring time-locked averaging across trials; Handy, 2005; Luck, 2014; Rugg & Allan, 2000). Both ERP and spectral analyses have been used to understand the temporal dynamics of memory processes. Several components of the ERP have been shown to index memory-specific computations, such as the negative slow wave, the contralateral delay activity or positive components, which typically emerge after ~250-400ms (Drew et al, 2006; Perez & Vogel, 2012; Voytek & Knight, 2010). In the spectral domain, memory signatures are mostly reflected in theta (4-8 Hz) or alpha oscillations (8-12 Hz; Klimesch et al. 1999; Kahana et al., 2001). As reviewed below, this enabled bridging findings from invasive recordings in rodents that showed a tight relationship between theta oscillation, neuronal firing and mnemonic content to human experiments. Spectral analyses have motivated several new investigations demonstrating that memory recall and mnemonic reactivations are modulated by a theta rhythm (Leszczynski et al., 2015; Kerren et al., 2018). Furthermore, EEG recordings that were obtained from lesion patients enabled dissecting the contributions of different cortical regions over time (Voytek & Knight, 2010a/b; Johnson et al., 2017).

Several recent theories have proposed that frequency-specific neuronal activity might subserve the selective routing of task-relevant information in the brain (Engel et al., 2001; Fries, 2015; Kohn et al., 2020; Siegel et al., 2012; Singer & Gray, 1995; Varela et al., 2001). EEG covers a wide-range of frequencies and hence, spectral decomposition of EEG data combined with source localization algorithms, connectivity analyses or decoding approaches now constitutes an efficient tool to study memory processes with high temporal and acceptable spatial resolution. Application of connectivity analyses have revealed that cortical memory networks are organized by theta rhythms (Sarnthein et al., 1998). Currently, high-density EEG recordings offer a comparable number of sensors as MEG. Each method has its advantages: EEG can be more readily obtained; MEG is less susceptible to electromyographic artifacts (Siems et al., 2016), but collectively both methods provide complementary views of the underlying physiology and have largely supported similar conclusions about the electrophysiological correlates of memory. See Chapter 4.5 "4.5. Oscillatory brain mechanisms for memory formation – Online and offline processes" which provides a detailed review of the literature on brain oscillations in memory research.

Polysomnography (PSG): Polysomnography has emerged as an important tool in the study of the effect of sleep on memory, including recent work on memory consolidation (see Chapter 6.9, "Sleep and Memory"). The PSG technique combines scalp EEG with electromyography (EMG), electrooculography (EOG) and electrocardiography (ECG) to detect distinct episodes of sleep. Non-REM (NREM) sleep can be staged from a few scalp EEG electrodes given the prominent occurrence of slow oscillations (SOs; < 1.25 Hz) and spindle oscillations (12-16 Hz, named according to their characteristic waveform). REM sleep is more difficult to detect from the EEG,

since it resembles patterns observed during wakefulness. Therefore, REM is typically defined based on the EOG (rapid saccade-like eye-movements) combined with a decrease in EMG amplitude (atonia). Recently it has become possible to stage REM sleep based on EEG features alone (Lendner et al., 2020) based on biophysical modeling of non-oscillatory population activity (Gao et al., 2017).

After the discovery of REM sleep (Aserinsky and Kleitman, 1953), sleep was initially conceptualized as an alternating sequence of 'inactive' non-REM and 'active' REM epochs. It had been assumed that memory consolidation primarily occurs during REM sleep given that signal characteristics mimicked wakefulness (Boyce et al., 2017). However over the last three decades, there is mounting evidence that memories are primarily re-activated and consolidated during NREM sleep (Buzsáki, 1996; Diekelmann & Born, 2010; Rasch & Born, 2013; Skelin et al., 2019) (see Chapter 6.9). The analysis of neuronal oscillations with scalp or intracranial EEG during NREM has revealed their key role in memory consolidation (Rasch & Born, 2013). In the active systems consolidation model, hierarchically nested sleep oscillations are thought to provide scaffolding for memory formation (Clemens et al., 2007; Rasch & Born, 2013; Staresina et al., 2015). Hippocampal ripples are associated with the reactivation and 'replay' of newly learned memories. Replay describes the phenomenon where a firing pattern that was present during encoding is reinstated during sleep (Buzsáki, 2015; Foster, 2017; Skelin et al., 2019; Todorova & Zugaro, 2018). Ripples do not occur in isolation but are nested, i.e. temporally coupled, into neocortical SOs and thalamocortical spindles through cross-frequency coupling (Clemens et al., 2007; Helfrich et al., 2019; Latchoumane et al., 2017; Maingret et al., 2016; Staresina et al., 2015). Cross-frequency coupling (Canolty & Knight, 2010) typically describes the observation that the phase of a slower frequency (e.g. SOs) modulates the amplitude of faster

events (e.g. spindles or ripples). Hence, these three cardinal sleep oscillations form a temporal hierarchy, which is thought to reflect endogenous temporal reference frames (i.e. windows-of-opportunity) for timed information transfer and consolidation. EEG remains the ideal tool to record oscillatory signatures during sleep. Novel data analysis strategies enable detailed analysis of coupling as well as the tracking of memory specific representations from PSG using decoding approaches (Helfrich et al., 2019; Schönauer et al., 2017; Zhang et al., 2018) and providing a more detailed picture than the sleep hypnogram.

Intracranial EEG (iEEG): iEEG is used in pharmaco-resistant epilepsy patients for seizure onset localization to guide the surgical resection of pathologic tissue (Parvizi & Kastner, 2018). To localize the seizure onset zone, neurosurgeons implant patients with stereo-tactically placed depth electrodes (sEEG) targeting medial temporal lobe and other deep structures and/or cortical grid/strip electrodes (Electrocorticography; ECoG; **Figure 2A**). Both methods record intracranial EEG with high spatial (sub-centimeter) and high temporal (sub-millisecond) resolution. Patients are typically monitored for 1-2 weeks in the hospital, during which anti-convulsive drugs are weaned off and cognitive testing is carried out in the patient room while continuous iEEG is being collected. It is best practice to remove channels in the seizure-onset zone or channels that exhibit epileptiform activity prior to data analysis (Ammanuel et al., 2020). Although patients with temporal lobe epilepsy and hippocampal sclerosis, which constitutes approximately two thirds of all cases, often exhibit baseline memory deficits, the qualitative features of their behavioral data closely resemble data from healthy adults (Parvizi & Kastner, 2018; Hill et al., 2020). Because epilepsy can potentially alter brain networks and physiology (Helmstaedter & Kurthen, 2001; Rao & Lowenstein, 2015), researchers use the patients as their own control,

comparing neural activity across experimental conditions within an individual. However, one cannot preclude that findings in this population would not generalize to healthy individuals. As such, findings should be corroborated through convergent evidence from other modalities (e.g., anatomical localization from fMRI studies; time-domain analyses from EEG/MEG).

iEEG provides the opportunity to assess the contribution of subcortical regions. Despite advances in source localization, non-invasive M/EEG are biased towards cortical activity given the spatial proximity. The erroneous notion that only cortical regions contribute to higher cognitive functions has reinforced a 'cortical myopia' (Parvizi, 2009) and a relative neglect of subcortical contributions to cognition. Intracranial EEG studies offer an ideal tool to evaluate contributions of subcortical regions to cognitive functions, revealing previously underappreciated associations, such as for instance the contribution of the hippocampus to visual attention (Slama et al., 2021).

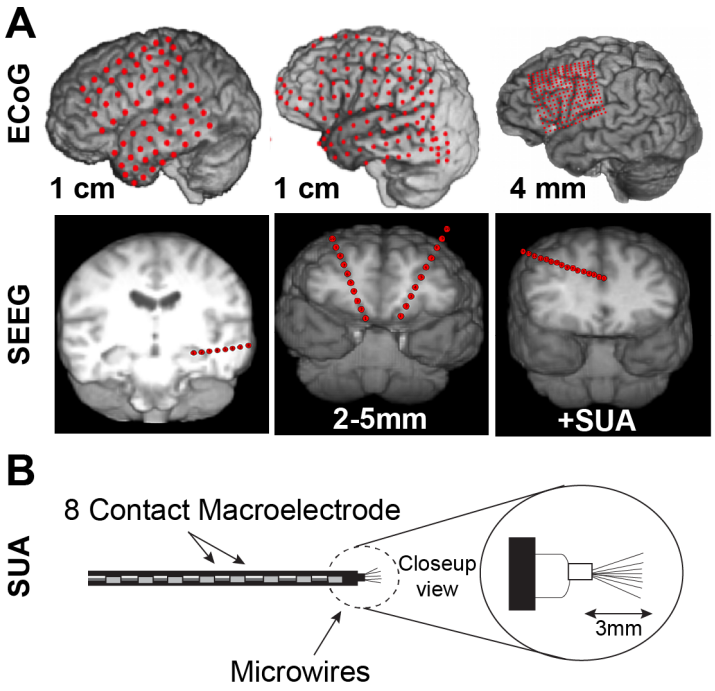
Despite its limitations, iEEG has substantially contributed to our understanding of memory processes and in particular the role of the MTL, which is difficult to study with high temporal resolution using non-invasive methods in humans. Key insights that intracranial EEG offered include a new appreciation of the role of oscillations in memory processes (Kahana et al., 2001). Specifically, intracranial recordings revealed that memory formation and recall is linked to patterns of regional synchronization and desynchronization (Hanslmayr et al., 2016; Johnson & Knight, 2015) and where different types of mnemonic information (e.g. spatial and temporal information) can be represented simultaneously in different frequency bands through multiplexing (Watrous et al., 2013). The analysis of high-frequency band activity (as surrogate of population spiking; Leszczynski et al. 2020; Ray & Maunsell, 2011) showed that this signal is also central to successful memory encoding and recall. Furthermore, typical implantation

schemes often involve bilateral, widely distributed electrode placement that can cover large-scale subcortical and cortical networks. Simultaneous recordings with high temporal resolution in the PFC and hippocampus have advanced our understanding of network mechanisms and the importance of directed information transfer for memory formation and recall (Herweg et al., 2021; Kragel et al., 2021; Long et al., 2017; Miller et al., 2013; Solomon et al., 2018; Solomon et al., 2019). The discovery of place and grid cells in rodent studies (O'Keefe and Dostrovsky, 1971; O'Keefe, 1976; Fyhn et al., 2004; Hafting et al. 2006; Sargolini et al. 2006), spawned interest if similar phenomena could be detected in the human brain. Several group started using 3D spatial navigation in a virtual reality environment to illuminate memory mechanisms in humans at both the single neuron and network level (Ekstrom et al., 2003; Jacobs et al., 2013;). Similar to rodent studies, it became evident that theta oscillations modulate spike timing and coordinate network activity during virtual navigation (Caplan et al., 2003; Kahana et al., 1999; Lega et al., 2012; Jacobs et al. 2010; Jacobs et al., 2013; Solomon et al., 2019). Overall, a highly comparable theta-mediated code was observed in humans. Several lines of inquiry indicate that the human brain might feature a larger variety of navigation cells (Kunz et al., 2021). A limitation is that all of these results were obtained from virtual navigation tasks and participants did not actually move around freely.

A novel treatment option for multi-focal seizure disorders, which cannot be treated by resective surgery, is the NeuroPace responsive neurostimulator (RNS) (Skarpaas & Morrell, 2009). RNS devices utilize chronic intracranial recordings, which detect epileptiform activity and automatically trigger electrical stimulation to disrupt the epileptic pattern. Recently, several studies have begun to take advantage of the opportunity to record from these chronically implanted electrodes and study memory processes that occur during real-world navigation

(Aghajan et al., 2017). This approach constitutes an exciting avenue to link rodent and human findings on the neurophysiological basis of spatial navigation. The first findings confirmed a key role of theta oscillations in coordinating network activity during real-world spatial navigation (Aghajan et al., 2017; Stangl et al, 2021).

Figure 2



Intracranial electrode placement. (A) Examples of intracranial EEG electrode placement: red dots depict individual electrode contacts. The first row highlights three examples of the commonly utilized ECoG grid electrodes with either 64 (left and center; 8x8 electrodes; 1cm inter-electrode spacing) or 256 electrodes (right; 16x16 electrodes; 4mm spacing). The second row illustrates stereo-tactically placed depth electrodes in the hippocampus (left), OFC (center) and cingulate cortex (right). Inter-electrode spacing and number of contacts is variable. Note that electrode contacts are present all throughout the shaft, allowing simultaneous recordings from subcortical and cortical regions, such as the temporal cortex (left) or PFC (center and right). (B) Depth electrodes might house additional wire bundles in their lumen, which allow recording of single and multi-unit activity at the tip of the depth electrode. Panel B is modified from Ad-Tech Medical Product Catalog Volume VII.

Single-unit activity (SUA): Several methods were introduced to obtain data at the single neuron level in humans (Fried et al., 2014; Ojemann et al., 2014; Rutishauser, 2019). Motivated by the discovery of place cells, grid cells and time cells as building blocks of the navigation system in rodents (O'Keefe and Dostrovsky, 1971; O'Keefe, 1976; Eichenbaum, 2014; Fyhn et al., 2004; Fyhn et al., 2008; Hafting et al. 2006; MacDonald et al., 2011; Sargolini et al. 2006), researchers were interested if similar principles guide human navigation. As reviewed above, at the network level comparable theta signatures emerged. In order to record single unit activity in the human brain, several approaches have been introduced. The typical approach is to insert an additional wire bundle (9 wires; 40μm diameter per wire; 0.9-1.3mm diameter of the sEEG electrode). through the lumen of a sEEG electrode, which records single unit activity at the tip of the depth electrode and is most commonly restricted to MTL and medial prefrontal regions (**Figure 2B**). The implantation of these additional wire bundles is generally considered to be safe (Carlson et al., 2018; Hefft et al., 2013). Similar data can be recorded during implantation of DBS (deep brain stimulation) electrodes in patients with Parkinson's disease from the sub-thalamic nucleus or the substantia nigra (Kamiński et al., 2018). However, DBS single-unit experiments need to be conducted in the operating room, a less than ideal scenario to study behavioral-physiologic processes yielding only 1-2 neurons per electrode. In addition, several groups have started to utilize microelectrode arrays (MEA; also termed the Utah array). These arrays are implanted into healthy cortex of tetraplegic patients to obtain high quality data to guide brain-computer-interfaces (Aflalo et al., 2015) or into tissue that is likely to be resected during epilepsy surgery (Cash & Hochberg; 2015; Truccolo et al; 2011; Vaz et al., 2020). Taken together, intracranial research with these recording methods provides novel opportunities to study memory function at the single unit level (Cash & Hochberg, 2015; Rutishauser, 2019).

Causal methods:

Lesion approach: Studying patients with brain lesions has provided important insights into human memory (Szczepanski & Knight, 2014; Vaidya et al., 2019). This approach requires a single focal lesion to the brain, which can result from various etiologies. However, lesion extent and etiology are often diverse and need to be taken into account when making inferences at the group level. Furthermore, as time since lesion onset increases, there is an increased probability that functional reorganization may occur. In addition, memory deficits could result from damage to passing fiber tracts, which further complicate functional localization. In lesion studies, one can typically use two types of controls: 1) patient behavioral or neural measures can be compared to those in a healthy, age-matched control cohort, or 2) neural measures in the healthy hemisphere can serve as the control in patients with unilateral brain lesions (Voytek et al, 2010; Vaidya et al., 2019).

Transcranial magnetic stimulation (TMS): TMS was developed in the late 1980s to non-invasively stimulate the brain and measure corticospinal activity during spinal surgery. This method takes advantage of the interaction of electric and magnetic fields: By transmitting an electric current through a coil, a magnetic field is induced, which in turn induces another electric field in the brain. The effect of TMS on the cortex depends on current intensity, coil shape and orientation, as well as on the current brain state. The typical application of TMS in cognitive neuroscience is the 'virtual lesion approach', where a single pulse or repetitive stimulation through a pulse train temporarily impairs the function of the stimulated cortex (Pascual-Leone et

al., 2000). TMS has undergone several technological improvements and innovations and is currently used in a variety of basic science (Ruff et al., 2009) and clinical (Rossi et al., 2009) applications.

Initially, TMS was used as a powerful tool to transiently perturb activity in a cortical area to study its contributions to a specific function (Pascual-Leone et al., 2000). Several stimulation protocols have been developed to either inhibit or excite the underlying cortex. In cognitive neuroscience, initial approaches stimulated with pulses, directly interfering with s cortical processing throughtout stimulation duration (Thut & Pascual-Leone, 2010; Censor & Cohen, 2011), or application os short burst of stimulation in the 3-8 Hz theta frequency band (referred to as theta-burst stimulation; Huang et al., 2005; Ziemann 2017), causing perturbation of cortical activity that outlasts the duration of stimulation. Subsequently, these protocols have been refined, leading to their widespread use as a tool to probe the causal function of a cortical region for a given cognitive process (Wassermann et al., 2008; Yeh & Rose, 2019; Bergmann et al., 2021; Pitcher et al., 2021).

Abundant evidence suggests that brain rhythms and population synchrony play an important role for cognition and several stimulation methods have been developed to modulate neural processing in a frequency-specific manner (Hanslmayr et al., 2014; Riddle et al., 2019; Thut, Veniero, et al., 2011). For example, frequency-tuned TMS is employed to selectively perturb or modulate function in a region or network (Hanslmayr et al., 2019). Synchronization of frequency-specific neural activity by stimulation (termed neural entrainment = directed synchronization through an external driving force) constitutes a powerful approach since it allows establishing the causal role of brain oscillations for memory and cognition (Herrmann et

al., 2016; Thut al., 2011). More recently, TMS has been used in a novel way to infer hidden or activity-silent network states (Rose et al., 2016) (see below section 3.2).

A shortcoming of TMS is that only cortical regions that are close to the skull, and not subcortical regions like the hippocampus, can be stimulated. However, several recent reports have demonstrated that the hippocampus can be influenced indirectly. For example, stimulation of parietal regions that are directly connected to the hippocampus impact hippocampal processing (Tambini et al., 2018; Wang et al., 2014). This intervention introduces long-lasting behavioral and network changes and constitutes a valuable tool to non-invasively modulate hippocampal dynamics in support of memory formation (Nilakantan et al., 2019).

Over the course of two decades, TMS has evolved significantly as a cognitive neuroscience tool, and is a prime example how one technique can be leveraged in several different ways to inform the study of human memory.

Transcranial electrical stimulation (tES): Researchers employ low intensity current stimulation (typically <= 2mA) to modulate neural activity below the threshold of action potentials (Nitsche & Paulus, 2000). In particular, the findings from Nitsche & Paulus suggested that passing a low intensity direct current (tDCS; transcranial direct current stimulation) through the brain slightly shifts the resting membrane potential up- or down-wards, resulting in an increase or decrease in cortical excitability. Subsequent research has demonstrated that this view is overly simplistic (Batsikadze et al., 2013; Giordano et al., 2017). However, multiple studies have shown that tDCS can modulate cognitive processes and improve memory function (Hill et al., 2016). Based on the idea that electrical activity patterns, such as theta (4-8 Hz) oscillations, are causally involved in

information processing (Kahana et al., 2001; Colgin 2013), studies have shown that alternating current stimulation (tACS) can produce a selective, frequency-specific modulation of neuronal activity (Helfrich et al., 2014; Polanía et al., 2012; Zaehle et al., 2010) This idea is rooted in non-linear systems theory and implies that neuronal oscillations can be entrained through an external driving force (Thut, Schyns, et al., 2011). Entrainment requires two narrow-band oscillators, one in the input stream and one that is being driven, which interact through directed synchronization (Pikovsky et al., 2003). Specifically the work by Polania et al. demonstrates that a selective modulation of neuronal population synchrony has direct effects on behavior, establishing a causal link between brain oscillations and behavior (Hanslmayr et al., 2019; Herrmann et al., 2016). tES methods are being refined and the precise mechanism-of-action is actively debated (Asamoah et al., 2019; Lafon et al., 2017; Vöröslakos et al., 2018).

Deep brain stimulation (DBS): DBS was originally developed as a therapeutic tool to treat deficits due to movement disorders, such as Parkinson's disease or essential tremor (Bronstein et al., 2011). More recently, it has also been explored as a tool to treat epilepsy and psychiatric disorders, such as obsessive-compulsive disorder (Mayberg et al., 2005). While several studies investigated the effects of chronic DBS of the hippocampus (Boëx et al., 2011; McLachlan et al., 2010; Miatton et al., 2011; Velasco et al., 2007), the effects on memory were often not significant, especially when compared to stimulation protocols that were only executed during specific phases of the task, such as the encoding or retrieval periods of memory tasks. A shortcoming of these studies was that an instantaneous read-out of the brain state was often not possible. The recently introduced NeuroPace RNS device combines continuous intracranial EEG recordings with the capacity for closed-loop stimulation (Suthana et al., 2018).

In the context of memory, another common target is the thalamus. For example, several studies have targeted the anterior nucleus (ANT) of the thalamus to stimulate ascending pathways to treat seizures in multi-focal pharmaco-resistant epilepsy (Peräkylä et al., 2017; Sweeney-Reed et al., 2014). These probes often also capture activity from adjacent nuclei, such as the medial dorsal (MD) nucleus, which exhibits prominent connections to the prefrontal cortex and might play a key role in coordinating large-scale memory networks (Schmitt et al., 2017). To date, stimulation-specific results on cognitive functions, such as memory, remain equivocal (Oh et al., 2012; Tröster et al., 2017).

Taken together, the reported effects of chronic DBS on memory processing and cognitive performance have been mixed. A major shortcoming of all studies was that stimulation was not tailored to specific phases of encoding, consolidation or retrieval, but rather often was determined the clinical protocol. In addition, electrodes for chronic DBS were not necessarily placed in central memory hubs. In order to test the more immediate effects of DBS, several groups utilized direct electrical stimulation during invasive EEG monitoring, which more regularly targets regions that are relevant for memory formation.

Direct electrical stimulation (DES): Direct electrical stimulation of the cortex can be carried out either during brain surgery or in the epilepsy monitoring unit (Selimbeyoglu & Parvizi, 2010). Typically, electric mapping of 'eloquent' cortex (i.e. speech and motor areas) is done prior to epilepsy or tumor resection surgery. During mapping, the cortex is stimulated with relatively high intensities in the range from 4-10 mA to *disrupt* neuronal processing. In cognitive experiments, the stimulation intensity is typically adjusted to values below the clinical mapping intensity (0.5 – 2mA) to *modulate* neuronal processing. Several studies take advantage of this

approach and combine cognitive testing with DES (Ezzyat et al., 2017; Fox et al., 2020; Jacobs et al., 2016; Kucewicz et al., 2018; Mankin & Fried, 2020; Suthana et al., 2012). In order to study human memory, electrical stimulation has been applied to medial temporal lobe structures, such as the entorhinal cortex or the hippocampus (Ezzyat et al., 2017; Fell et al., 2013; Hansen et al., 2018; Khan et al., 2019; Mankin & Fried, 2020), but modulatory effects have also been observed after stimulation of other structures of the limbic system, such as the fornix (J. P. Miller et al., 2015), lateral temporal cortex (Ezzyat et al., 2017; Ezzyat et al., 2018) or the amygdala (Inman et al., 2018). To date, experimental findings have been equivocal, with some studies demonstrating an improvement of memory (Suthana et al., 2012), while others observed detrimental effects of stimulation (Jacobs et al., 2016). In an attempt to reconcile divergent findings, it has been argued that the effects of stimulation might be dependent on the performance at baseline, the current network state, the stimulation location (e.g. white or gray matter) or the precise task phase (Titiz et al., 2017). In order to overcome some of these limitations (Suthana et al., 2018), several recent studies utilized a closed-loop approach, where stimulation is individually tailored according to the instantaneous electrophysiological brain state (Ezzyat et al., 2018). For example, the above mentioned RNS device delivers stimulation whenever an epileptic discharge is detected. In the context of the study of memory, stimulation could be tailored to theta features (power or phase) to modulate ongoing activity in a close-loop fashion. Jointly, all of these studies suggest that stimulation constitutes a promising avenue to modulate human memory processes in-vivo (Hanslmayr et al., 2019), however, additional work is necessary to clarify where, when and how stimulation should be delivered to reliably boost memory performance or to alleviate memory decline in patients suffering from memory loss.

Microstimulation: In contrast to regular DBS or DES, microstimulation is delivered through delivery of low-current electrical stimulation via microelectrodes (typical diameter <100μm vs. >1mm in DBS/DES). A major advantage of this approach is that spatial targeting is easier, since the effects of stimulation are spatially more confined. To date, stimulation protocols have often been inspired by protocols that were pioneered in animals and have been proven to induce long-term plasticity, such as theta-burst stimulation (Histed et al., 2013). Critically, and unlike DBS, stimulation intensities are in the range of physiologic level currents (~150μA). Therefore, it is believed that microstimulation mimics the natural conditions that drive plasticity in the brain. To date, most evidence stems from animal experiments (Fetsch et al., 2014; Logothetis et al., 2010). More recently, several groups pioneered this approach in humans (Schmidt et al., 1996) and demonstrated that microstimulation can influence reinforcement learning (Ramayya et al., 2014) as well as memory specificity for novel items (Titiz et al., 2017). Given that microstimulation requires the additional implantation of microwires in addition to macroelectrodes, this approach has not been widely adopted to date.

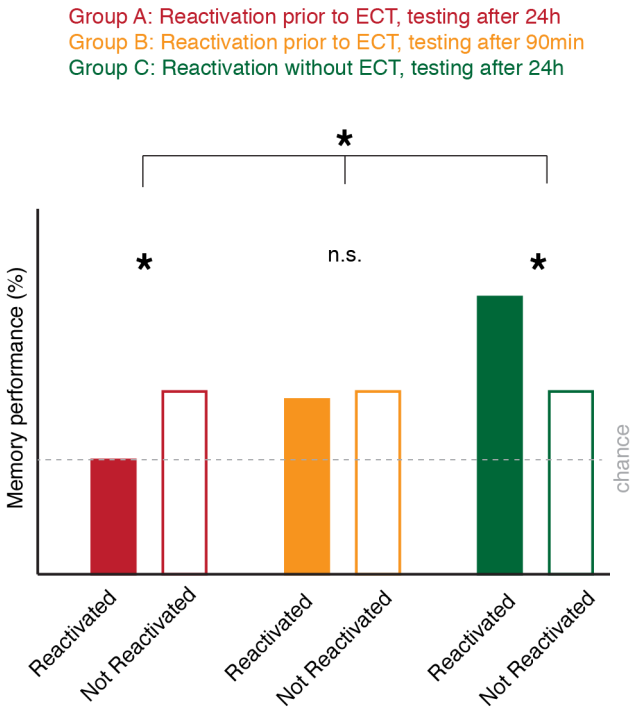
Electroconvulsive therapy: ECT is a clinical procedure designed to induce a seizure by applying high currents through the skull, typically under general anesthesia. It is a powerful tool to treat pharmaco-resistant depression (Pagnin et al., 2004)., but may also have long-term detrimental effects on cognition and memory (Sackeim, 2000). In a series of experiments, Squire and colleagues studied the effects of ECT on memory performance (Squire, 1977; Squire et al., 1976, 1984). They observed that ECT induced a profound, but temporally-graded amnesia for previous testing sessions. Moreover, they showed that skill learning was preserved. As the interval between learning and ECT is increased, the resulting retrograde amnesia diminished (Squire et

al., 1976). Thus, after ECT the patient may describe a past event accurately but be unable to report when the event occurred. More, recently this seminal work has been extended from studies of consolidation to studies focusing on reactivation (information is brought from an inactive to an active state) and reconsolidation (stabilization of the memory trace).

In the past, ECT had been used as a tool to study temporally graded retrograde amnesia (Cohen & Squire, 1981; Devanand et al., 1995; Squire et al., 1975; Squire et al., 1976). Using this approach, it had been observed that ECT affects recent memories more profoundly than past memories. For example, memories that have been acquired years ago were less affected. This finding supported the notion that over time memories undergo a consolidation process (Frankland & Bontempi, 2005; Winour & Moscovitch, 2011).

Recent work has combined behavioral testing prior to ECT to study memory consolidation demonstrating that ECT selectively disrupts memory consolidation of recently reactivated information (Kroes et al., 2014; **Figure 3**), thus, further supporting the reconsolidation hypothesis. Typically, subjects experience a benefit from the reactivation (relative to control group C who did not receive ECT). However, if memories reactivate just prior to ECT (group A), subjects do not consolidate these memories. ECT did not affect non-reactivated memories (Kroes et al., 2014). Thus, ECT can be utilized as a tool to study the temporal evolution of memories. Depending on the experimental design, studies using ECT as an intervention may unravel the time-course of memory formation, consolidation and reactivation. A critical shortcoming is that the method can only be applied in subjects who are prescribed ECT as a treatment for medication refractory depression, which in itself is already often associated with impaired memory (See Chapter 9.8, "Memory, Depression and Anxiety").

Figure 3



Probing memory consolidation following ECT. Patients with pharmacoresistant depression who underwent ECT where randomly assigned to one of three groups (Only groups A and B received ECT). Prior to ECT one of two previously learned memory associations was reactivated (solid vs. empty bars). As expected, memory recall was improved for reactivated items over non-reactivated items – but only in group C, which did not receive ECT. In group A, which received ECT after reactivation but was tested after a 24h delay, performance dropped to chance and below the non-reactivated items. Critically, group B was tested after 90mins and reactivated memories remained unaffected. This finding suggests a specific disruption of a time-dependent consolidation process by ECT. Schematic according to the findings as described by (Kroes et al., 2014).

2.2 Integration of distinct neuroscientific methods

Each of the methods presented above has strengths and limitations, such as the restricted temporal resolution of fMRI, the limited spatial resolution of M/EEG or the uncertain

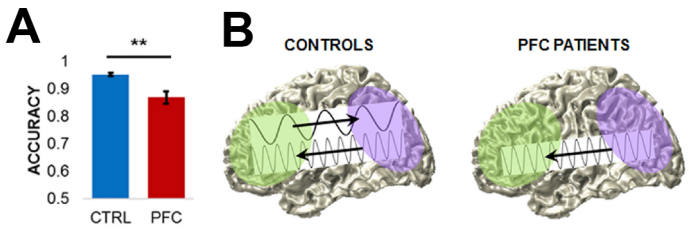
physiological efficacy of non-invasive brain stimulation. Furthermore, patient studies involve inferences affected by variable lesions, damaged neural tissue, and brain reorganization (D'Esposito, 2010).

To address these limitations, researchers can combine data from multiple neuroscientific methods, leading to enhanced group-level inference. One successful approach is combining fMRI and EEG (Mulert & Lemieux, 2009). Albeit correlative, combining localization information from fMRI with timing information from EEG revealed that frequency-specific activity in the theta/alpha/beta-range provides a functional mechanism to bind MTL and PFC networks during memory encoding, maintenance or recall (Hanslmayr et al. 2011; Herweg et al., 2016; Scheeringa et al., 2008). While iEEG has a high spatiotemporal resolution, data can only be obtained from a few brain regions of the network and does not allow whole-brain coverage (but see Solomon et al., 2017 for an exception). Simultaneous iEEG and fMRI or M/EEG can potentially overcome this limitation to some degree (Dalal et al., 2013). Likewise, simultaneous fMRI or EEG, and transcranial magnetic stimulation suffer from concurrent artifacts that are introduced into the recordings by the magnetic pulses (Ruff et al., 2009; Thut et al., 2011). Lesion studies are most powerful if behavioral testing is combined with either EEG or fMRI to directly link their deficits to physiological recordings (Vaidya et al., 2019).

This multimodal approach has recently been used to investigate network-dependent working memory function in patients with prefrontal damage. Patients and healthy controls were presented with two visual items, which they had to remember over a few seconds and then had to make a judgement about the second pair with respect to the first one. As expected, healthy subjects (controls) performed better than lesion patients, however, even the PFC lesion patients performed significantly above chance (~85 % correct; chance level: 50%; **Figure 4**) (Johnson et

al., 2017). Simultaneous scalp EEG and network connectivity analysis revealed that prefrontal damage attenuated a prefrontal-dependent theta network (4-8 Hz), but left a parieto-occipital alpha/beta network (10-30 Hz) intact. Given that the patients still demonstrated some degree of task proficiency, Johnson et al concluded that working memory relied on the parieto-occipital alpha/beta network more so than the prefrontal theta network. Although technically challenging, integrating information across neuroscientific methods confers significant advantages over single-method studies.

Figure 4



Bidirectional Frontoparietal Connectivity Supports Working Memory. (A) Behavioral results. Memory recall performance was better in healthy controls than in PFC lesion patients, but patients performed well above chance level (0.5). (B) Schematic illustration of two distinct bidirectional systems supporting WM. While the bottom-up (posterior to PFC; purple to green) system in the alpha-/beta-range remained intact in PFC patients, the top-down PFC-dependent delta-/theta-system was attenuated in patients. Given that the PFC lesion patients still demonstrated task proficiency, it suggests that the bottom-up system may be sufficient for WM, whereas the top-down system is not, perhaps only exerting moderate modulatory influences on performance.

3. Analytical approaches to understanding human memory

In this section we consider how the methodological tools described above have advanced our current understanding of memory function. Specifically, we provide illustrative examples of how these techniques have been employed to study human memory. As the succeeding chapters of the handbook provide detailed reviews in specific areas our goal here is to provide a general overview of how novel analysis strategies have shaped our view of human memory.

3.1 Activation vs. Representation

Traditionally, researchers inferred memory processing from activity patterns (Curtis & D'Esposito, 2003; Fuster & Alexander, 2001; Goldman-Rakic et al., 1995; Miller & Cohen, 2001; Pasternak & Greenlee, 2005). For example, memory-encoding processes have been linked to differences in BOLD signals, event-related potentials or neuronal firing rates by contrasting activity patterns for remembered and forgotten items. However, several studies focusing on neocortical association areas have reported that subjects can hold information in their mind without any discernible cortical activation (Stokes, 2015); thus, raising the question of whether mnemonic representations might be encoded in large-scale synaptic dependent connectivity patterns.

To address this question, several novel analytical tools have been employed, ranging from information-theoretical tools (Quian Quiroga & Panzeri, 2009) to decoding analyses (Grootswagers et al., 2017) and representational similarity analysis (Kriegeskorte et al., 2008). These tools are all geared towards defining spatiotemporal pattern$_S$ (representation) of memory

processes. In this framework, representation does not require activation per se (as defined as an activity increase relative to baseline, which is often used in univariate analyses), but relates the overall observed pattern across space (channels or voxels) and/or across time to the mnemonic information (Kriegeskorte et al., 2008). By aggregating information from multiple brain signals, this approach offers greater statistical power for detecting subtle differences. Furthermore, representation could not only entail the joint activity pattern at multiple spatial locations, but this framework also allows quantifying connectivity patterns of distributed networks (Eichenbaum, 2017; Yuste, 2015). Below we outline several recent concepts that go beyond the activation-based framework and highlight how using novel analytical tools has shaped our understanding of memory processing.

3.2 New tools lead to new concepts about human memory

Activity-silent coding: Activity-silent coding proposes that mnemonic information can be maintained through rapid shifts in synaptic weights within local neural networks (Stokes, 2015; Wolff et al., 2017). In this framework, the encoding of new information is proposed to trigger short-term plasticity, contributing to changes in synaptic weights. However, despite the technical progress, it is currently not possible to image neuronal networks at the level of synapses in humans, hence, Wolff et al. devised an indirect approach to detect short-term plasticity. This framework implies that shifting in synaptic weights is not associated with a change in the overall activity pattern, as observed using fMRI or M/EEG (Stokes, 2015). In other words, the activity-silent coding framework posits that mnemonic information can be encoded in the brain, but in a

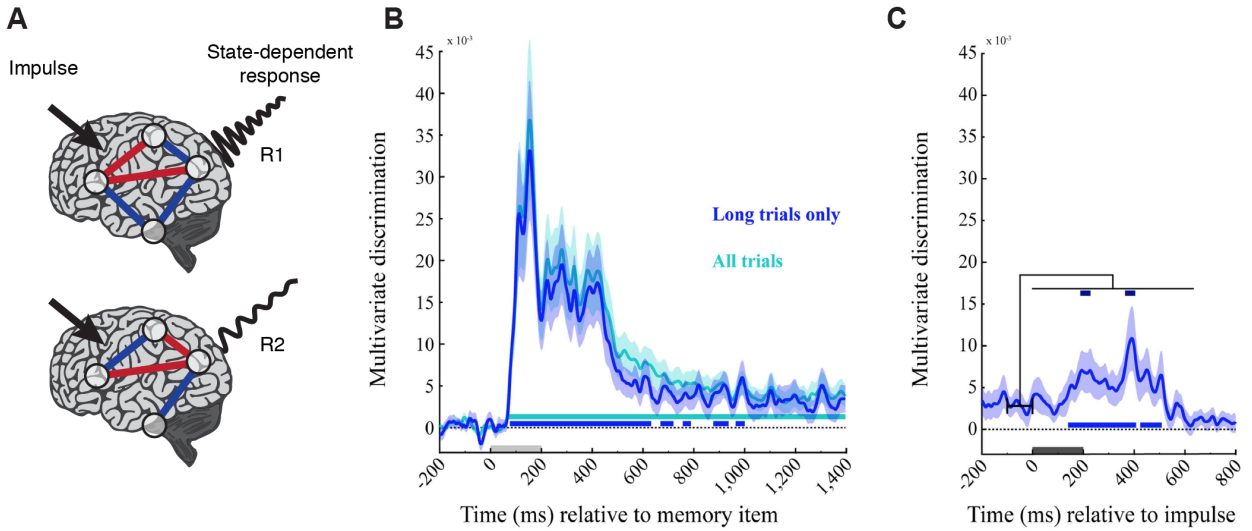
latent state, which cannot directly be quantified by studying overt activity patterns. In a novel approach that was utilized to infer these hidden network states, Wolff and colleagues devised a 'pinging' approach (illustrated in **Figure 5A**). The method can be compared to a ship sonar that searches the ocean floor using repetitive sound waves. They provided repeated sensory input (reflecting the sonar signal) into the brain (Wolff et al., 2017) and then recorded the neural response (reflecting the sonar echo). Wolff et al. reasoned that the response will be modulated as a function of the underlying brain state (similar to a modulation of the sonar echo frequency once an object is detected on the ocean floor). In their experiment, Wolff et al. approach utilizes a non-informative high-contrast visual stimulus (analogous to emitted sound waves), which is presented during the delay period when the information is kept in memory to probe the network (Wolff et al., 2017). Critically, the network response (the echo) differed according to the item held in memory, and indexing the underlying cortical state. This approach offers the opportunity to indirectly infer the content of memories from hidden or latent cortical states conceptually similar to assessing the unobservable bottom of the sea (Sreenivasan & D'Esposito, 2019).

The key interpretation of this finding is that memories might be stored in 'activity-silent' states and that mnemonic information might be encoded in connectivity profiles at the level of synaptic weights in distributed cell assemblies.

In a related study, Rose et al (2016) used a TMS pulse instead of a visual stimulus, but the key idea remained the same: Probing the network and reading out its response (echo), which should differ as a function of the contents of memory. By keeping the external pulse the same, one can infer the current network state (and indirectly the mnemonic content) from the evoked response. Hence, the authors concluded that depending on the content of memory a transient perturbation

will lead to a distinct network response, hence, offering an instantaneous read-out of items held in memory (**Figure 5**).

Figure 5



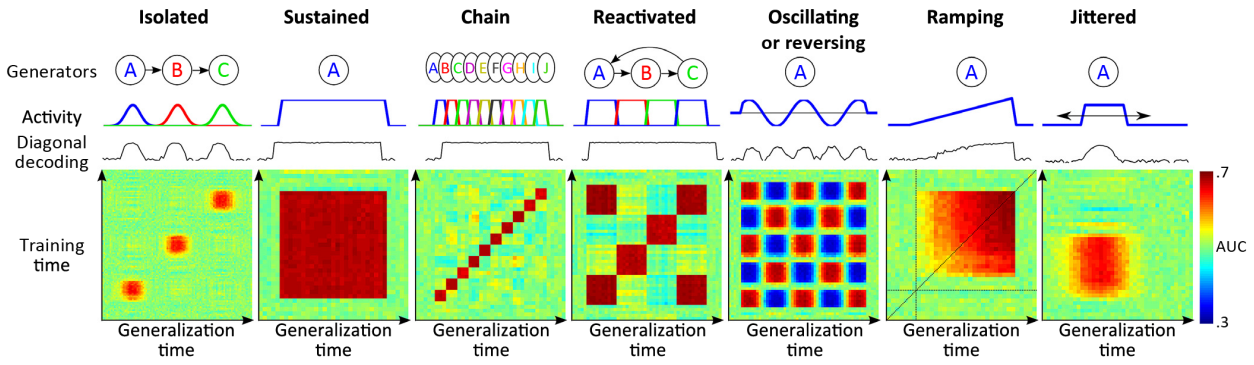
How to infer latent states through perturbation. (A) Schematic of 'pinging' approach: One constant impulse into a network may generate differential responses (R1 and R2) depending on the instantaneous network configuration. The precise network depends on the content of memory. (B) Data from a sensory perturbation experiment: There is significant decodability after stimulus presentation, which however, returns to baseline after approximately 800ms. However, the subjects still hold on to the same mnemonic information, thus, raising the question why this information is not accessible by decoding strategies. (C) If an uninformative, high contrast visual stimulus is presented during the time of non-decodability to perturb the system, then one can successfully 'read-out' the content of working memory. This finding indicates that one can decode latent states from a network perturbation suggesting that the precise network pattern contains content-specific information. Figure panel B and C are reproduced with permission (Wolff et al., 2015) under the Creative Commons Attribution (CC BY) license.

Static vs. dynamic coding: The concept of static and dynamic coding directly emerged from the multivariate analyses of brain data. Multivariate analyses reflect a diverse set of algorithms that can detect unique patterns in high-dimensional datasets, hence often termed multivariate pattern

classifiers (MVPC) or analysis (MVPA) (Haxby, 2012; Hebart & Baker, 2018). These algorithms are sensitive to subtle differences. The major difference to univariate analyses is that pattern classifiers can distinguish differences in the configuration (e.g. across space: pattern across electrodes in response to stimulus A differs to pattern in response to stimulus B). This approach is commonly utilized when no difference in the average response is apparent (ERP to stimulus A = ERP to stimulus B) (Grootswagers et al., 2017). Multivariate analyses of neural data can adjudicate between different neural coding strategies that would appear identical under univariate metrics. MVPC analyses always require partitioning the data into a 'training' dataset (where the classifier knows which trial belongs to which condition) and a 'test' dataset to which the classifier is being applied. When compared to the ground truth, one can infer if the classifier correctly classified more trials than expected by chance.

In the context of memory research, MVPCs have been used to test whether the neural pattern that is present during encoding of a memory at an earlier time remains the same at a later time (King & Dehaene, 2014). If the classifier still performs significantly better than chance, then one can infer that the overall pattern, which discriminates both conditions, has not changed, which would indicate 'stable' coding. In contrast, if classifier performance drops to chance, then one can conclude that the representation must have changed, i.e. the neural coding scheme is 'dynamic' (**Figure 6**).

Figure 6

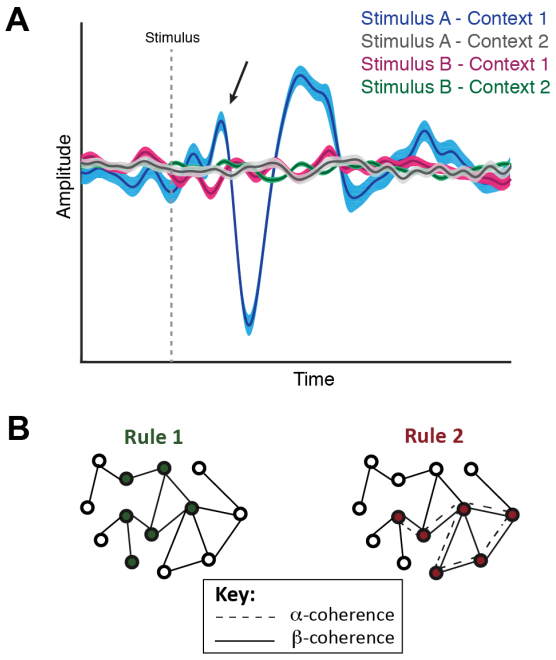


Principles of using classification to assess static vs. dynamic neural codes. The temporal cross-generalization matrix (lower row) depicts the relationship of training and testing time. If a temporal code is specific for a time-point, then one will only observe significant decoding if the decoder is trained and tested on the same time points (panel 1 and 3). This is visible by significant 'on-diagonal' decoding. This time-varying coding scheme is also termed dynamic coding. In contrast, if one observes significant 'off-diagonal' decoding (panel 2 and 6), then training on one time point is generalizable to a different testing time point, suggesting that the neural code remained the same (or static). Several additional scenarios are possible, which reflect variants of the static and dynamic frameworks (e.g. temporally selective reactivation: panel 4; oscillatory modulation of coding: panel 5; temporal blurring since cognitive processes are not perfectly aligned across trials: panel 7).

Mixed selectivity: Mixed selectivity describes the context-dependent coding of information (**Figure 7**) . This concept suggests that neurons or neuronal populations in association cortex can engage in several processes. To date, mixed selectivity has only been observed at the single neuron level, but might be an important concept for understanding human memory (Ekstrom et al., 2003; Rigotti et al., 2013). For example, Rigotti et al. showed that the same neuron fires in response to a stimulus only during passive recognition (context 1) but not during active recall (context 2). It has been argued that mixed selectivity might reflect the fact that different network states subserve distinct cognitive functions and that neurons are recruited into a transient cell assembly through neural synchrony (Saxena & Cunningham, 2019). In other words, neurons can

be engaged in different networks, which are formed through neural synchrony (Fries, 2015; Siegel et al., 2012). In the context of memory, this idea implies that a neuron could code more than one piece of information. The coding scheme might be dynamic and the precise encoded information might depend on whether a particular neuron is recruited into a transient coalition of synchronously active neurons. This notion contradicts the classic neuron doctrine that viewed neurons as passive feature detectors, but rather suggests that a single neuron can contribute to a wide-range of behaviors (Yuste, 2015).

Figure 7



Mixed selectivity and network connectivity. (A) Context-dependent activity: The black arrow indicates that there is only a significant activation in response to stimulus A in context 1, but not in context 2. There is no response to stimulus B in either context. (B) Different neurons could become engaged in different operations depending on network coherence, where multiple canonical computations can be mapped on the same network depending on the frequency-specific network organization. Panel B summarizes data from Buschman et al. (Buschman et al., 2012).

Sustained vs. transient activity: Temporally sustained activity patterns have been reported across a range of studies, including single unit recordings in primates, BOLD fMRI, EEG event-related potentials and oscillatory power (Sreenivasan & D'Esposito, 2019). Neural responses in fMRI and EEG are typically quantified as trial averages given the relatively low signal-to-noise ratio of the method. In the case of EEG, averaging in the time-domain reveals prominent phase-locked components, also termed event-related potentials, such as the P300 (Sutton et al., 1965) or the contingent negative variation (Walter et al., 1964), which have successfully been assessed in a variety of cognitive experiments (Helfrich & Knight, 2019a; Polich, 2007; Soltani & Knight, 2000). While this approach is powerful in isolating phase-locked components, it averages out non-phase-locked components, which often have been considered to reflect 'noise'. In addition, non-phase-locked components typically occur in beta/gamma frequencies, thus, averaging in the time-domain is implicitly biased towards lower frequencies (David et al., 2006; Tallon-Baudry & Bertrand, 1999). In the context of memory, sustained responses have been observed in the trial-averaged spike traces recorded in monkey prefrontal cortex (Fuster & Alexander, 1971; Goldman-Rakic, 1995), which has often been conceptualized as the neurophysiological signature that reflects 'keeping a memory in mind' (E. K. Miller & Cohen, 2001). More recent evidence suggested that this sustained response may in part be an artifact of averaging across many trials with slightly different spike timing (Lundqvist et al., 2016). When inspecting single trials, researchers noticed that single trials rarely exhibit sustained responses. When focusing on spikes in the memory-delay interval, Lundqvist et al. noticed spike-locked beta and gamma signatures, which did not show in the trial-averaged spectrograms due to their variable nature (Lundqvist et al., 2016). These activity bursts were only discernible at the single-trial level, which led to a series of empirical (Lundqvist et al., 2018; Spaak et al., 2017) as well theoretical (Earl K. Miller

et al., 2018; M. Stokes & Spaak, 2016) investigations. Moving from trial-averaged to single-trial analyses has led to a reconceptualization of delay activity in memory processes (Sreenivasan & D'Esposito, 2019), thus, requiring a different tool set that does not solely rely on time-averaged data to understand how mnemonic information is represented.

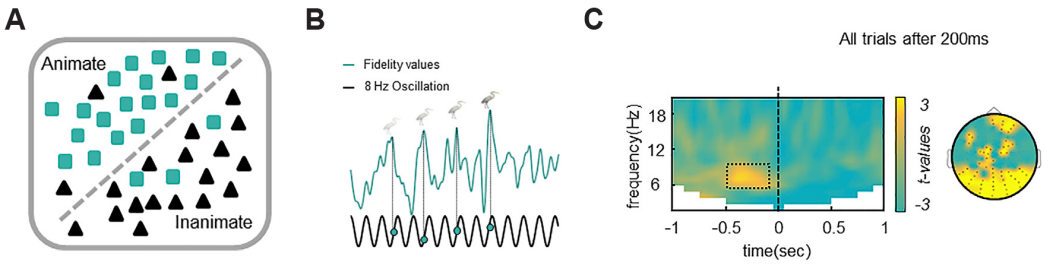
Rhythmic reactivation of memories: Much EEG activity is non-oscillatory in nature and characterized by a power law (He et al., 2010; Lendner et al., 2020; Miller et al., 2009; Voytek & Knight, 2015). Cognitive operations, and specifically memory tasks, are often hallmarked by the emergence of 3-8 Hz theta oscillations (Anderson et al., 2010; Herweg et al., 2021; Lisman & Jensen, 2013; Johnson et al., 2017). It had previously been observed that the phase of theta oscillations structures neuronal firing and providing temporal reference frames for different spikes (Colgin, 2013), Furthermore, theta has been suggested to provide a mechanism of how spatial information can be represented in a temporal code as discussed earlier, where theta groups e.g. place cell activity and mapping out spatial trajectories over time (Skaggs et al., 1996; Wilson & McNaughton, 1993).

In addition, the emergence of theta is often accompanied by gamma oscillations. Theta and gamma power correlate with successfully memory encoding, but might constitute two distinct entities with distinct origins (Sederberg et al., 2003; Fellner et al., 2019). However, it had been observed that theta phase also structures gamma-band activity (Lisman & Jensen, 2013) in a process called theta-gamma cross-frequency coupling (CFC; Mormann et al., 2005; Canolty et al., 2006; Canolty & Knight, 2010; Tort et al., 2010) where theta phase predicts the instantaneous gamma-band amplitude. Theta-gamma CFC has been shown to support memory formation in the hippocampus (Axmacher et al., 2010; Lega et al., 2012; Staudigl & Hanslmayr, 2013; Tort et al.,

2009). More recently, it had been shown that theta-gamma CFC segregates the representation of different pieces of mnemonic information within a single oscillatory cycle (Bahramisharif et al., 2018). Contemporary theories posit that theta-gamma CFC may represent a potential biophysical mechanism that explains how information can be encoded, even after a single stimulus presentation (Hanslmayr et al., 2016; Lisman & Jensen, 2013). Different phases of neuronal oscillations might provide temporal reference frames to separate representations of multiple items held in memory, and thereby reduce conflict and increase the fidelity of mnemonic information (Leszczyński et al., 2015; Lisman & Jensen, 2013; Siegel et al., 2009).

In addition to structuring the content of memory, a related hypothesis suggests that encoding and retrieval might be linked to distinct theta phases providing temporal segregation (Hasselmo, 2005; Hasselmo et al., 2002). Additional support for this hypothesis has recently been provided by Kerren and colleagues (2018) who showed that only certain theta phases trigger the reactivation of mnemonic information (**Figure 8**). They utilized multivariate-pattern classifiers to track mnemonic representations and observed that decoding peaks were systematically predicted by theta phase alignment, thus, indicating that tight link between theta phase and memory reactivation (Hasselmo, 2005). Taken together, multiple lines of research indicate that memory encoding, maintenance and retrieval critically relies on theta-rhythmic processes in the human brain.

Figure 8



Information reactivation is nested in theta oscillations. (A) A pattern classifier was used to decode object categories from working memory in a time-resolved manner. (B) Schematic of the relationship of decoding output (classifier confidence) and the underlying theta oscillation. Note the classification peaks coincide with distinct phases of the theta oscillation. (C) Realignment of the data to the classification peaks (time point 0) demonstrates significant theta phase concentration prior to reactivation and category information peaks. These results highlight a tight relationship between information reactivation and theta oscillations.

4. The network structure of human memory

Mnemonic representations can be found at every level of the cortical hierarchy, which highlights the need to understand both the contribution of individual nodes of a network, such as the prefrontal cortex or hippocampus, as well the brain-wide network structure (Johnson & Knight, 2015). To study networks, data must be collected simultaneously from extended brain regions, which is not easily accomplished with single-unit physiological techniques, but is possible with tools such as fMRI, M/EEG or intracranial EEG (Solomon et al., 2017), which will be the focus of this section.

4.1 How to assess network connectivity

Connectivity between brain regions can be classified into either structural or functional connections. Here, we focus on functional connectivity, which allows tracking of dynamic changes over a short period of time during memory formation, consolidation and reactivation. Functional connectivity can further be grouped into undirected and directed interactions (Bastos & Schoffelen, 2015).

Because the hemodynamic response function is relatively slow, connectivity derived from fMRI BOLD signal primarily captures very slow fluctuations over time. Correlation of FMRI time series amplitudes provides a metric of undirected functional connectivity, which can also be spectrally decomposed (~0.01 – 1 Hz). Time-domain Granger causality allows assessing directed connectivity, i.e. node A driving node B or vice versa (Seth et al., 2015). Information-theoretical metrics such as mutual information and transfer entropy constitute a model-free approach to study network connectivity (Quian Quiroga & Panzeri, 2009).

M/EEG data has a rich temporal structure and captures a wide range of frequencies (~0.1 – 100 Hz). Several seminal theories emphasized that synchronization of band-limited oscillatory activity might subserve information transfer in large-scale networks; hence, it is best practice to spectrally decompose the signal prior to connectivity analysis (Fries, 2015; Pesaran et al., 2018; Siegel et al., 2012). Furthermore, spectral power of M/EEG signals declines with the reciprocal of frequency (the so-called 1/f^a power spectrum) implying that low frequencies dominate correlations in the time domain (Miller et al., 2009). Most methods for undirected connectivity are derived from the coherence metric, which considers both phase- and amplitude-based connectivity (Engel et al., 2013). Several variations have been introduced, either suppressing the

amplitude contribution (e.g. phase-locking value, phase-lag index, pairwise phase consistency) or ignoring the phase contribution (e.g. amplitude-envelope correlations). Effects of electrical volume spread in the cortical tissue can be attenuated using bipolar referencing in intracranial EEG (Solomon et al., 2019) recordings combined with connectivity metrics (i.e. imaginary coherence, weighted phase-lag index, orthogonalized amplitude-envelope correlations) that suppress zero-phase lag interactions to extract true inter-areal interactions (Bastos & Schoffelen, 2015). In addition, there is a growing literature on connectivity as measured directly from intracranial electrodes (Foster et al., 2013; Johnson et al., 2018; Solomon et al., 2017). Findings obtained from intracranial recordings, for example, assessed information flow within in the hippocampus formation (Axmacher et al., 2008; Fell et al., 2006), between amygdala and hippocampus (Inman et al., 2018; Zheng et al., 2017), between prefrontal and association cortices (Johnson et al., 2018; Watrous et al., 2013) and between mesial and lateral temporal cortices (Vaz et al., 2019), confirming and extending rodent studies (Buzsáki, 2015; Johnson & Knight, 2015). Collectively, these studies provide the necessary scientific context and validation to relate intra- and extracranial studies and interpret connectivity effects.

However, the directionality of the interactions cannot be inferred from most connectivity metrics, i.e. whether a given region is the sender or the receiver (information flow from region A to B or vice versa). Several methods, such as Granger causality (Seth et al., 2015), phase-slope index (Nolte et al., 2008) or transfer entropy (Lobier et al., 2014), cab quantify directional interactions (see Phan et al., 2019 for a critical discussion a possible solution using multivariate modeling). These methods typically quantify if the history of a signal in region A can predict the future of the signal in region B. If so, then one can infer information flow from A to B. Despite featuring the word 'causality', these metrics do not actually measure 'causal' interactions, but only enable

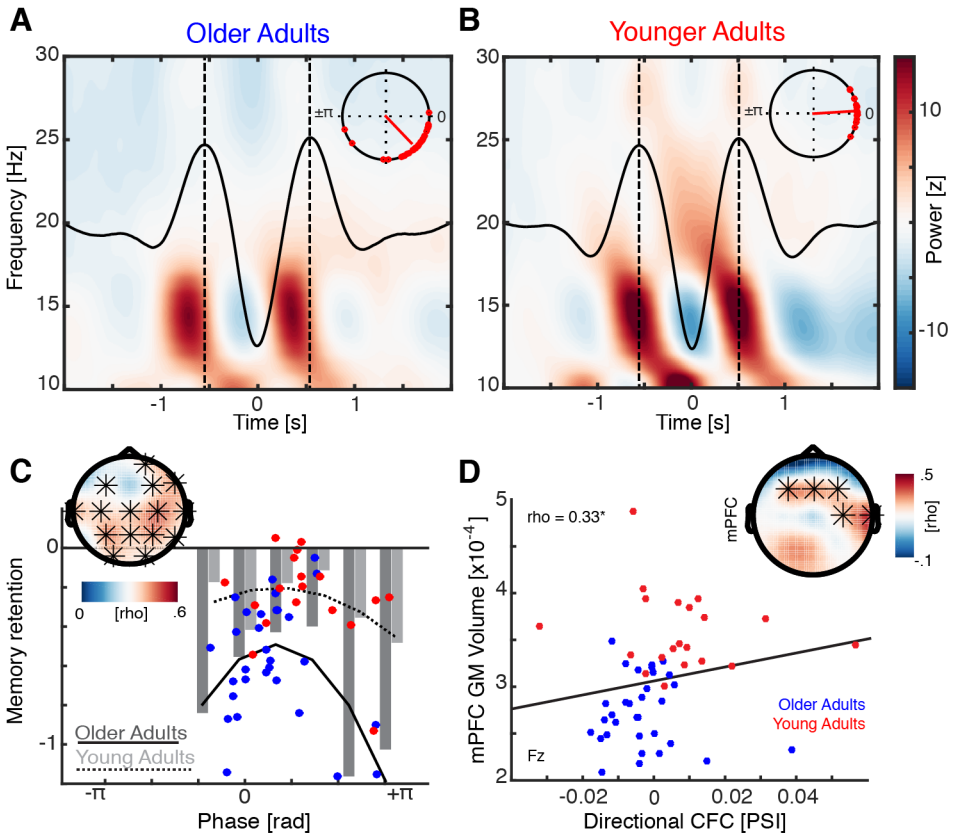
model-based inferences about the directionality of these effects (i.e. flow from A to B is more likely than B to A). Notably, similar metrics can be used to study interactions across temporal scales (coordination of activity with different frequencies), i.e. cross-frequency coupling (CFC) (Aru et al., 2015; Canolty et al., 2006). CFC is as a powerful concept to explain information transfer across different temporal scales and has been shown to play a pivotal role for both short-term (theta-gamma coupling) (Lisman & Jensen, 2013) as well a long-term memory formation (SO-spindle coupling during sleep) (Helfrich et al., 2019; Staresina et al., 2015). There are a multitude of connectivity metrics but it is still unclear how different modes of interaction (e.g. amplitude- or phase-based) or the metrics derived from different imaging modalities are related (Hall et al., 2014). Several studies report prominent differences in phase- and amplitude-based connectivity metrics, suggesting that phase- and amplitude-based connectivity might reflect distinct connectivity modes to support multiplexing in large-scale networks (Engel et al., 2013). Comparative fMRI and M/EEG connectivity studies have not found a clear relationship between the metrics derived from these two methods (Hipp & Siegel, 2015). Hence, there is not a direct transfer function between metrics derived from fMRI and M/EEG.

4.2 The importance of timing in large-scale networks

Memory formation likely depends on the precise timing between different cortical regions. Information transfer in large-scale networks depends on the coordinated interplay between specialized cortical nodes (Helfrich & Knight, 2016; Siegel et al., 2012). Over the last two decades, several lines of research have indicated that phase-aligned activity across different network nodes is necessary for information transfer in large-scale networks (Engel et al., 2001; Fries, 2015; Siegel et al., 2012). Critically, a shift of the precise delay or a temporal dispersion

(Kohn et al., 2020) impacts the flow of information in neuronal networks (Johnson et al., 2018), and thus, is detrimental for memory formation or recall (Hanslmayr et al., 2016; Polanía et al., 2012). More recently, the same mechanism was implicated in coordinating different temporal scales (Griffiths, Parish, et al., 2019; Siebenhühner et al., 2016). Helfrich et al. compared overnight memory formation in younger (~20 years) and older adults (~70 years) who performed a hippocampus-dependent memory task (Helfrich et al., 2018). As expected, younger participants, on average, exhibited a higher recall performance the next morning than older subjects. However, brain activity during sleep as measure by EEG was remarkably similar: Both groups exhibited numerous oscillatory events (distinct events detected in the time domain), such as slow oscillations (SOs; < 1.25 Hz) and spindle (~12-16 Hz) oscillations, indicating that these cardinal sleep oscillations alone did not predict differences in memory function. When probing the fine-tuned temporal interaction between SOs and spindles by cross-frequency coupling, Helfrich et al. found that spindles arrived too early during the SO cycle in older adults (**Figure 9**). Crucially, we observed the same pattern in both groups: Memory performance declined with more spindles missing the optimal coupling phase. Deviations as small as 50-100ms predicted impaired memory performance and correlated with increased gray matter atrophy in the medial prefrontal cortex in older adults, where SOs are generated. Taken together, when considering timing in a large-scale network, both timed information transfer across space, as well as across temporal scales, are necessary for successful memory formation and recall.

Figure 9



Impaired slow-wave spindle coupling predicts memory deficits. (A) SO trough-locked time-frequency representation (TFR) reveals elevated spindle power just prior to the SO peaks (dashed lines) in older adults. The inset highlights the average SO-spindle coupling phase across 32 older adults. (B) SO trough-locked TFR demonstrates that states of high spindle power coincide with SO peaks in younger adults. Same conventions as in panel A. (C) The precise SO-spindle coupling phase predicts overnight memory retention. In both groups, less forgetting was associated with more optimal coupling closer to the SO up-state (around 0°). (D) The strength of the directional influence of the SO phase on spindle power correlates with grey matter (GM) volume in the mPFC suggesting that aging impairs the temporal coordination of SOs and spindles and impairs memory performance.

5. Conclusions and future directions

This chapter surveyed neuroscientific methods to study human memory. In addition to approaches in animals and healthy humans, we have also provided several examples of neuroscientific approaches originally designed for clinical applications that have been be utilized to elucidate the principles and concepts of human memory. Our understanding of human memory will continue to be shaped by the development of novel data acquisition and analysis approaches. In particular, multimodal investigations that bridge single-unit and population-based techniques will illuminate the functional basis of memory in humans. Furthermore, we foresee that information-theoretical analysis tools that link experimental and computational approaches will provide the necessary means to integrate evidence across different modalities into a coherent model of human memory (Griffiths, Mayhew, et al., 2019), while refined stimulation protocols will help to establish causality (Hanslmayr et al., 2019).

Scientists have long searched for mnemonic engrams, i.e. the long-lasting physiologic changes in the brain caused by a stimulus, which reflect a memory. To date, it remains unclear which level of observation is needed to detect an engram. In humans, memory traces have been observed with fMRI BOLD signal, EEG, local field potentials and single neuron activity. In rodents, engrams are often directly associated with the replay of a particular neuronal firing sequence (Foster, 2017). The quest for the identification of mnemonic engrams not only requires a methodological choice, but also awareness that certain methods can only test a certain framework: Does one expect that the single neuron is the unit of cognition (Barlow, 1953) as famously suggested by Horace Barlow, where neurons are thought to serve as feature detectors, which are precisely tuned to a very specific stimulus property (Eichenbaum, 2017; Hubel &

Wiesel, 1962), or does one conceptualize the engram according to Donald Hebb (Hebb, 1949) who highlighted the importance of cell assemblies for cognition? While Barlow's idea is very much in line with the concept of neuronal replay, where the specific reactivation of a certain cell is thought to consolidate memories, the majority of available methods in humans favor Hebb's idea that cell assemblies and analyzing activity at the population level will better inform behavior (Saxena & Cunningham, 2019). Hebb's concept is advantageous to explain cognitive flexibility, since many memories can be mapped on a given neural circuit (Fusi et al., 2016; Rigotti et al., 2013; Saxena & Cunningham, 2019). Note that the finding that single neuron activity correlates with behavior is not inconsistent with the view that cell assemblies code information (Rutishauser, 2019). An important question is how we conceptualize cell assemblies. Are cell assemblies composed of specialized neurons that code for e.g. objects, space, time or do cell assemblies reflect transient coalitions of 'multi-tasking' neurons that can code for more than one concept. This idea has previously been referred to as mixed selectivity (Rigotti et al, 2013; Fusi et al., 2016), which might provide a powerful conceptual framework for future investigations into memory functions.

Hence, a multimodal approach across several spatiotemporal scales is needed not only to test existing theoretical frameworks, but also to bridge and integrate these diverse findings into a coherent framework that can explain memory formation in humans. We predict that methods that have been pioneered in rodents will be used in humans to illuminate the inner workings of memory. Recent developments include NeuroPixel recordings from hundreds to thousands of neurons (Paulk et al., 2021; Steinmetz et al., 2021; Stringer et al., 2019), imaging the brain using light (i.e. near infrared spectroscopy; Obrig & Villringer, 2003), modulating the brain using focused ultrasound (Folloni et al., 2019; Verhagen et al., 2019) or possibly even optogenetics

(Deisseroth; 2011; Boyden et al., 2015). Irrespective of the method, we are convinced that closed-loop interventions hold great potential to establish causality and offer a potential treatment of memory disorders in the future. Despite a surge of new methods, it is evident that not a single method can solve how human memory systems operate. It will be critical to integrate methods to bridge spatial (units, cell assemblies, local populations, network-wide activity) and temporal scales (μs to years).

Taken together, the field of human memory has benefited from a wealth of new methods that have helped to pinpoint the 'when' and 'where' of memory formation, while the main challenge for future studies is the understanding of 'how' memories are (trans-)formed.

Acknowledgements

This work was supported by the German Research Foundation (DFG HE8329/2-1 to R.F.H.), NIH grants MH63901 and MH11173 (to M.D.) and NIH NS21135 to (R.T.K.).