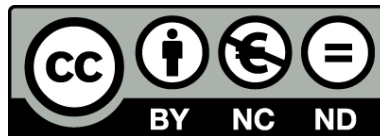




UNIVERSITAT DE
BARCELONA

Synaptic and circuit mechanisms of working memory and their dysfunction in anti-NMDA receptor encephalitis and schizophrenia

Heike Stein



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement- NoComercial – SenseObraDerivada 4.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento - NoComercial – SinObraDerivada 4.0. España de Creative Commons.**

This doctoral thesis is licensed under the **Creative Commons Attribution-NonCommercial-NoDerivs 4.0. Spain License.**



UNIVERSITAT DE
BARCELONA

Doctoral Program in Biomedicine

Neuroscience

Neurophysiology and computation in cortical systems

**Synaptic and circuit mechanisms of working memory
and their dysfunction in anti-NMDA receptor encephalitis and
schizophrenia**

Heike Stein

Barcelona

October 2020

A handwritten signature in blue ink, appearing to read "H. Stein".

Heike Stein
(PhD candidate)

A handwritten signature in blue ink, appearing to read "Albert Compte".

Albert Compte
(thesis supervisor)

A handwritten signature in blue ink, appearing to read "Josep Dalmau".

Josep Dalmau
(thesis tutor)

Acknowledgements

The past four years in Barcelona have had an immense impact on me personally, and on who I have become as a scientist. Many people have contributed to this transformation, and I am grateful for the patience, knowledge, friendship and trust I have experienced during these years.

First and foremost, I want to thank Albert Compte, and acknowledge the centrality of his intellectual guidance for my work and my scientific maturation. It was a privilege to have you as a supervisor and mentor. I can hardly imagine entering science in a better way than by joining your lab.

I want to thank Josep Dalmau: Being a giant in your own field of neurology, you have had the vision and trust to embark on a cognition- and computation-centered journey. Together with Myrna Rosenfeld, you have been a source of support during these years.

This thesis is founded on the work of many, and it was inspiring, and at times challenging, to work in a truly interdisciplinary project. A few people deserve special acknowledgements for closely accompanying the experiments in their most critical moments: Diego, Adriá and Alba, who have worked hard for this project's success.

Rosselló 149, -1, if windowless, was an amazing learning environment, thanks to the incessant, passionate discussions inside and outside the lab: Thanks to Jaime de la Rocha, for not ever letting the lab's passion for science fade, and for being vocal, engaged, and always interested. Thanks to (the extended) Neurobiología Teórica: Dani Linares, Lejla, João, Ainhoa, David, Genis, for whom there was no topic too profane or too complex to be debated to its last detail, and to Neurobiología Disfuncional and Circuitos Corticales: Jordi, Tiffany, Pablo, Yerko, Balma, Lluís, Manuel, Dani Duque, Carles, Rafa.

Barcelona for a neuroscientist is a one-of-a-kind place in Europe, thanks to the dense, vivid BARCCSYN network. Many of you have become close friends over the years: Alex Hyafil, Jose, Gabriela and Klaus, Ramón, the two Marias; again, Genis, Lejla, Ainhoa, and my first friend in Barcelona, Lucia. Thank you for taking me to Croata, your homes and terraces, the beach, concerts, and to beautiful places in Cataluña, Spain & Portugal. I sincerely wish and trust to see you in many more places as our friendship continues. João, you have had a special role in all this; I learned from you for life, and I am lucky that our paths crossed in -1.

I want to acknowledge the people who know me the longest: First of all, my parents, who have given me the privilege and encouragement to live my life the way I want to, and my grandparents: Your achievements are immense, as you reinvented yourselves in a foreign place. My brother, who has become a true ally in adulthood. A last thank you to two persons who safeguarded my sanity for many years now, Anna and Nina: You are by no means “bitter women”, rather the opposite.

Finally, I wouldn't have the same perspective and motivation to become a scientist without two long summers spent in Shanghai and Woods Hole. Sharing knowledge, doubts, and friendship with the next generation of world-class neuroscientists felt like winning the lottery.

Which brings me to Claudia Clopath, Lluís Fuentemilla, Gabriela Mochol, Klaus Wimmer, and Josep Marco-Pallarés: Thank you for critically reading and discussing with me the work that emerged from these past four years.

Table of Contents

Acknowledgements	4
Table of Contents	7
Table of Abbreviations	9
Chapter 1: Introduction	11
1.1 Working memory: How does the brain maintain information in the absence of inputs?	12
Working memory as a fundamental cognitive function	12
Cellular mechanisms of working memory	15
From cellular to network mechanisms	17
How does persistent delay firing translate to human neuroimaging and EEG?	20
What is the role of synaptic plasticity in working memory?	21
1.2 Errors and systematic biases in working memory	23
Forgetting	23
Diffusion	24
Systematic biases	25
1.3 The role of the NMDAR for working memory	29
The NMDAR contributes to persistent activity	30
Synaptic plasticity on multiple timescales depends on the NMDAR	32
Schizophrenia	35
Anti-NMDAR encephalitis	37
Chapter 2: Goals	41
Chapter 3: Results	43
Declaration of impact and author contributions	45
3.1 Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory	47
3.2 Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia	61
3.3 Working memory codes are reactivated between trials in healthy controls, but not in anti-NMDAR encephalitis and schizophrenia	73

Chapter 4: Summary of Results	93
Chapter 5: Discussion	95
5.1 How do these findings impact our understanding of working memory?	96
5.2 Does NMDAR dysfunction not affect working memory maintenance, after all?	100
5.3 Which aspects of altered cognition does a disruption in serial dependence in anti-NMDAR encephalitis and schizophrenia reflect?	104
Chapter 6: Conclusions	109
Appendix	111
A1 Supplementary material for Chapter 3.1	113
A2 Supplementary material for Chapter 3.2	131
A3 Supplementary material for Chapter 3.3	155
A4 Ten simple rules for modern psychophysics	161
Bibliography	177

Table of Abbreviations

AMPA	α -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor
BOLD	blood oxygen level dependent
Ca ²⁺	calcium
CSF	cerebrospinal fluid
E/I balance	excitation/inhibition balance
EEG	electroencephalography
EPSC	excitatory postsynaptic current
EPSP	excitatory postsynaptic potential
fMRI	functional magnetic resonance imaging
GABAR	gamma-aminobutyric acid receptor
LTP	long-term potentiation
Mg ²⁺	magnesium
Na ⁺	sodium
NMDAR	N-methyl-D-aspartate receptor
ODR	Oculomotor delayed-response task
PFC, dlPFC	prefrontal cortex, dorsolateral prefrontal cortex
PPC	posterior parietal cortex
PSTH	peristimulus time histogram
STP	short-term plasticity
V1	primary visual cortex

Chapter 1

Introduction

1.1 Working memory: How does the brain maintain information in the absence of inputs?

Consider the most basic building blocks of animal behavior: Moving between different places, maybe with the goal of finding sources of food; in humans, having a conversation or planning the tasks of one's day. In each of these situations, animals have to think ahead and plan the necessary actions to reach their goal. More fundamentally, animals have to represent goals throughout a continuous span of time, in order to concatenate different actions that will lead to their accomplishment. A fox that is chasing prey, for example, has to remember that it has seen a rabbit in order to follow it through the woods. Similarly, a presenter at a conference has to remember the beginning of their sentence in order to finish it coherently. Even though these situations are vastly different in many respects, they share the feature that without an ongoing representation of past information, successful behavior would not be possible. One way or another, the brain has to achieve continuity of information through time, especially in the absence of relevant inputs.

In Chapter 1.1, I will present the solutions that neuroscience has proposed for this problem. First, I will define working memory from a cognitive perspective and introduce classical experimental tasks that are used to measure properties of working memory across species and sensory systems (note, however, that the focus of this thesis lies on visual working memory). Then, I will review electrophysiology findings from the last 50 years that are foundational to our understanding of how neural circuits implement working memory. Finally, I will discuss how mechanisms of short-lived synaptic plasticity could emerge from, and contribute to, working memory maintenance.

Working memory as a fundamental cognitive function

The set of brain functions that achieve a temporally continuous representation of past experiences is called memory. Depending on a number of factors, but mostly based on the timescale over which information is maintained, psychologists, cognitive scientists and

neuroscientists distinguish between short-term or *working memory*¹, and *long-term memory*. This distinction has first become evident in patients with cortical lesions, in whom information maintenance on either shorter or alternatively longer timescales was affected, depending on the location of the lesion (Scoville and Milner 1957; Shallice and Warrington 1970). Still, these functions are by no means independent, as working memory can be influenced by long-term contents, and working memory might play a role in the initial formation of long-term memories. For now, I will provide a definition of working memory as a brain function that retains information in an active state over short periods of time (typically hundreds of milliseconds to several seconds), thereby allowing for concurrent maintenance and processing of the stored information (Baddeley 1992). Measures of working memory are strongly correlated with different aspects of higher-order cognition, including reasoning, learning, planning, language comprehension, and general cognitive ability (Conway et al. 2003), pointing to its core contribution to most aspects of cognitive performance.

In parallel to early studies of working memory in humans (Atkinson and Shiffrin 1968; Baddeley 1986), its study in animal models has allowed to make significant advances beyond purely cognitive theories by measuring brain activity on the level of single neurons while animals perform working memory tasks (Fuster and Alexander 1971; Kojima and Goldman-Rakic 1982). Some of the groundbreaking findings in primates, described in the next section, became fundamental to the understanding of how the brain maintains information in the absence of inputs. Therefore, in this thesis I will mostly focus on tasks that employ low levels of abstraction and can be used in both human and animal experiments (as compared to e.g. verbal working memory tasks, which limit us to human subjects and therefore rarely allow for direct assessments of single-cell activity).

Working memory across species can be measured in different sensory domains (visual, tactile, and auditory (Pasternak and Greenlee 2005), but also olfactory, which is a natural choice for rodent models (Liu et al. 2014)). While important differences exist regarding task structures, all share the basic building blocks of working memory *encoding* (information is presented in the form of a stimulus), *maintenance* (the stimulus is removed, but the information must be remembered over several seconds), and *recall* (the stimulus identity must be reported: either by reproducing its value, adjusting a second stimulus to match the

¹ where working memory, in contrast to short-term memory, is not just limited to storing information, but implies a more active maintenance (Baddeley 1992).

remembered one, choosing the previously shown stimulus from a set of stimuli, or a simple, “yes/no” binary identification of a probe). A specific type of working memory task, the *delayed-response task* (Hunter 1913), is shown in Figure 1. Other than locations, modern versions of the task can test the retention of various other visual features, such as orientations, spatial frequency or movement direction, different auditory features, such as tone frequency, loudness, etc. Popular variations of this task, with the goal of making it more challenging and testing the limits of working memory, are the sequential or simultaneous presentation of several to-be-remembered stimuli, the introduction of distracting stimuli during the delay, or the introduction of contextual information that interferes with the presented stimuli (e.g. by introducing stimulus correlations between trials).

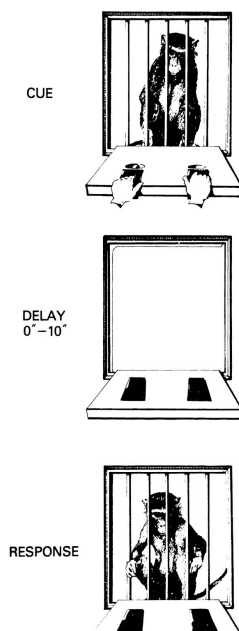


Figure 1. Example of an “analog” visuo-spatial working memory or delayed-response task. In this early version of the task, the experimenter shows the animal a peanut (*cue*) on one of two locations (left or right). Then, an opaque screen is lowered for several seconds, during which the animal needs to remember the presented location (*delay*). Finally, the blind is removed again, and the animal can choose at which side it remembers the peanut to be (*response*). In more modern versions of this task, the stimulus is presented on a computer screen, and the to-be-remembered stimulus value can lie on a continuum (e.g., an angular location instead of a left/right decision). Moreover, other stimulus dimensions than location can be used, such as gratings with different contrasts, spatial frequency or movement speed for visual stimuli, tone frequency or loudness for auditory tasks, etc. Figure from Goldman-Rakic (1987) with permission from Wiley publishing group.

Depending on the type of stimulus and the recall modality, working memory can be quantified in terms of the somewhat interdependent terms of *capacity* (How many presented stimuli are recalled?), *accuracy* (Is the recalled stimulus correctly identified or reported?), and *precision* (How precisely does the recalled or reproduced stimulus value match the encoded one?). As we shall see in the Chapter 1.2, multiple sources of errors exist and can lead to forgetting or imprecision of working memory representations in the brain. To understand these phenomena in a mechanistic framework that takes the neural circuit as its starting point, it is important to first discuss the neural correlates that have been identified in animals performing working memory tasks.

Cellular mechanisms of working memory

In the early seventies, single-cell activity during delayed-response tasks was measured for the first time in primates (Fuster and Alexander 1971; Kubota and Niki 1971). Informed by lesion studies in humans and monkeys that resulted in working memory deficits (Brutkowski 1965), these researchers chose to record in the prefrontal cortex (PFC). Fuster and Alexander (1971) found that more than half of the 110 recorded prefrontal cells increased their spiking activity transiently or during a prolonged period of time during stimulus presentation, delay, or both, as compared to a pre-stimulus baseline. Intriguingly, many neurons showed changes in activity that were sustained throughout the whole delay², sometimes during more than 60 s.

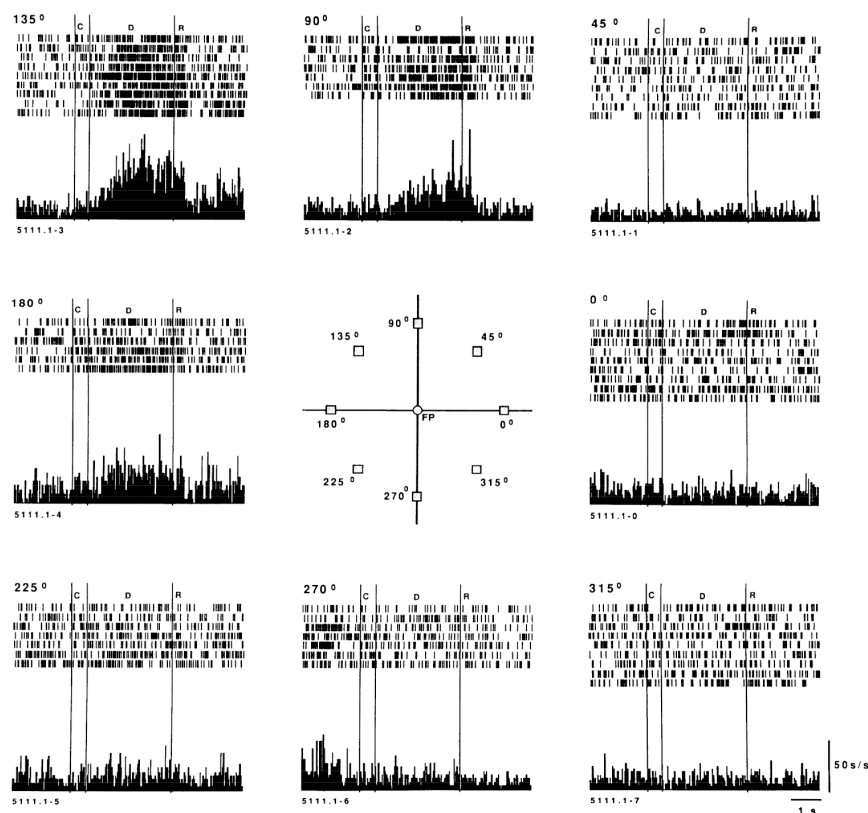


Figure 2. One neuron's responses during stimulus presentation (C), delay (D), and saccadic response (R) in the ODR (see text below). Stimuli in each trial were presented at any of eight locations (middle). On the outer plots, spike times (vertical lines) for an example neuron are plotted, each subplot corresponding to a group of trials (rows) with the same stimulus location. Below, a peristimulus time histogram (PSTH) shows cumulative spike counts over several trials with the same stimulus location. Note that activity is smoothly modulated by stimulus location, showing more similar responses for adjacent stimuli. A suppression of activity for opposite locations can be seen in other example neurons that are not shown here. Figure from Funahashi et al. (1989) with permission from The American Physiological Society.

² The proportion of persistently active neurons in PFC was estimated as 33% in a follow-up study two years after these initial findings (Fuster 1973).

The finding of prefrontal *persistent activity* was replicated often in the years to follow (for an extensive review, see Leavitt et al. 2017). Importantly, Niki and Watanabe (1974; 1976) showed soon after that prefrontal activity not only spanned the whole delay period, but at the same time encoded the identity of the memorized stimulus: Independently of the upcoming motor response direction, a considerable share of cells exposed consistent side preferences, i.e. they increased their activity more strongly if the stimulus had appeared on one side rather than the other. However, classical monkey electrophysiology setups like the one shown in Figure 1 made it hard to precisely control for motor and postural confounds, stimulus timing, and to introduce more elaborate stimulus protocols. To overcome these limitations, Funahashi et al. (1989) designed a computer-controlled oculomotor delayed-response task (ODR) that allowed to present visual stimuli across a wide part of the monkey's visual field, and that measured eye movements or *saccades* as responses. Their most important result, shown in Figure 2, summarizes what has since then been widely accepted as the neural substrate of (visuospatial) working memory: Location-tuned, persistent activity in PFC. Around 70% of neurons with persistent activity exposed location preferences. The authors showed that these preferences arise from smooth location tuning similar to neural tuning in primary visual areas (Hubel and Wiesel 1959), and thereby coined the term *memory fields*.

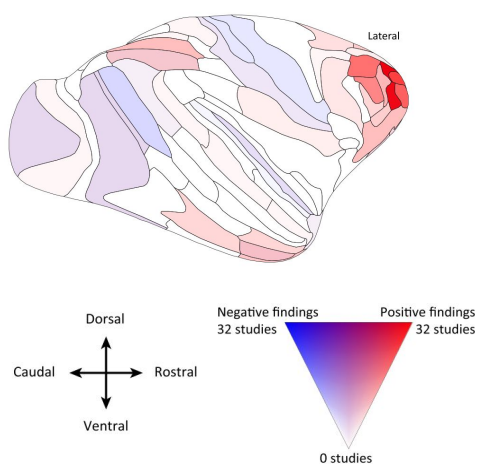


Figure 3. Areas of the primate brain in which stimulus-specific persistent activity during visual working memory is consistently found. Color hue indicates the number of positive (red) or negative (blue) results, and saturation levels the number of studies conducted for each area (the more saturated, the more studies, out of max. 32 reviewed studies). Other than PFC (rightmost), regions with persistent activity include primarily inferotemporal (center bottom) and parietal (center top) regions. While early sensory regions are probably mostly involved in memory encoding, association cortices with persistent activity are suited for domain-general working memory storage and manipulation (Xu 2017; Panichello and Buschman 2020). Figure adapted from Leavitt et al. (2017) with permission from Elsevier.

It is important to note that persistent activity during working memory delays has not only been found in PFC (despite being the most unequivocally identified region). A recent

review of over 90 studies in primates by Leavitt et al. (2017) reports evidence for stimulus-selective, sustained activity in several higher-order, associative cortices, including prefrontal, parietal, and inferotemporal regions. In contrast, evidence for sustained activity in early sensory regions is rather sparse (Figure 3), and is potentially explained by top-down feedback from PFC (Xu 2017).

From cellular to network mechanisms

After a cellular substrate of working memory maintenance had been identified, questions arose on how a neuron could produce constant output over such long timescales. One possibility is that specific membrane properties of single neurons could lead to single-neuron *bistability*, so that a neuron can fire with both low baseline firing rates or sustained high firing rates in the absence of external input. The sustained firing state is achieved through extended periods of excitability that cause prolonged, so-called *plateau potentials*, and exceed the time of stimulation by several hundred milliseconds or, in rare cases, substantially longer periods (Marder et al. 1996; Zylberberg and Strowbridge 2017). While evidence for single-cell persistent activity has been found in PFC (Haj-Dahmane and Andrade 1998) and intrinsic, non-task-related timescales of PFC neuronal firing are in fact longer than in most other cortical areas (Murray, Bernacchia, et al. 2014), single-cell persistent firing in most neurons still barely spans sufficiently long periods for successful memory maintenance (usually during delays of up to ~ 20 s; Inagaki et al. 2019).

An easier solution to the problem of persistent firing becomes evident when one considers the PFC as a network: Its neurons receive inputs not only through feedforward pathways from sensory brain regions, but also from recurrent connections that form feedback loops (Lorente De Nó 1938; Hebb 1949). While feedforward input ceases soon after external stimuli are removed, *recurrent connections* feed output from a PFC neuron back to itself (after further processing through other neurons of the same or a different circuit³). Indeed, neurons in PFC (similar to other higher-order cortices) communicate with other PFC neurons through horizontal connections within the same layer (González-Burgos

³ Candidate regions that might form recurrent loops with PFC are those that show persistent activity themselves, primarily inferotemporal and parietal regions (for a review, see ref. (Riley and Constantinidis 2015)). A rather unexplored hypothesis of the origin of PFC persistent activity during working memory lies in recurrent connections of the PFC with the thalamus, basal ganglia, and, as pointed out by more recent evidence, cerebellum. Early on, persistent activity during delayed-response tasks has been found in the thalamus (Fuster and Alexander 1973), and recently, Gao et al. (2018) have shown that frontal cortex persistent activity depends on inputs from deep cerebellar nuclei, which also exhibited persistent activity in a delayed-response task.

et al. 2000; Kritzer and Goldman-Rakic 1995), as illustrated in Figure 4. Moreover, correlated delay responses in adjacent neurons (Rao et al. 1999) suggests a circuit structure that favors connections between neurons with similar memory fields.

The advances in the neuroanatomical and electrophysiological understanding of the PFC made in the 70s, 80s and 90s inspired biophysical models of the prefrontal microcircuit as a system that maintains stimulus-specific working memory representations autonomously through recurrent excitatory and inhibitory connections (Goldman-Rakic 1995). The principal phenomenon in this family of models (Compte et al. 2000; Amit and Brunel 1997; Lim and Goldman 2013) is *reverberation* of activity through time, which in turn depends on sufficiently strong recurrent excitatory connections (Wang 2001). When stimulus-specific inputs are fed into these networks, a transient increase of activity in location-tuned neurons occurs. These neurons send outputs to similarly tuned excitatory neurons, while inhibitory connections suppress activity in neurons that are more dissimilar in tuning. Inhibition is achieved through a pool of inhibitory interneurons, which is smaller than the excitatory pool (usually with a ratio of ~ 1:4). The balance between excitation and inhibition ensures that activity in the microcircuit is *stable* both in the absence of memoranda (the state of spontaneous activity, in which neurons usually fire with rates of ~1-4 spikes/second), and in a state of persistent delay firing (in which stimulus-specific neurons show firing rates of ~20-100 spikes/second). Importantly, *network bistability* implies a non-linear relation between input strength and neural firing, where the network switches abruptly between spontaneous and persistent firing as inputs exceed a certain threshold. Figure 5 depicts activity in a bump-attractor network for working memory (Compte et al. 2000) that simulates PFC activity recorded in Funahashi et al. (1989).

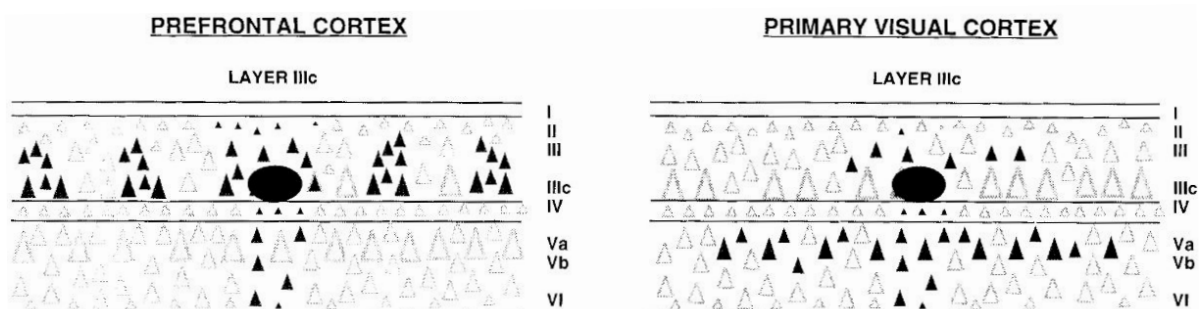


Figure 4. Retrograde labeling in monkey PFC vs. primary visual cortex (V1) identifies the neurons that provide input (triangles) to a cell (oval) located in a specific layer, here layer 3c. In contrast to V1, layer 3 PFC neurons receive extensive inputs from other neurons of layer 2, 3 and layer 3c. Similar patterns of connectivity are observed in layer 5 (not shown). Figure from Kritzer and Goldman-Rakic (1995) with permission from John Wiley and Sons.

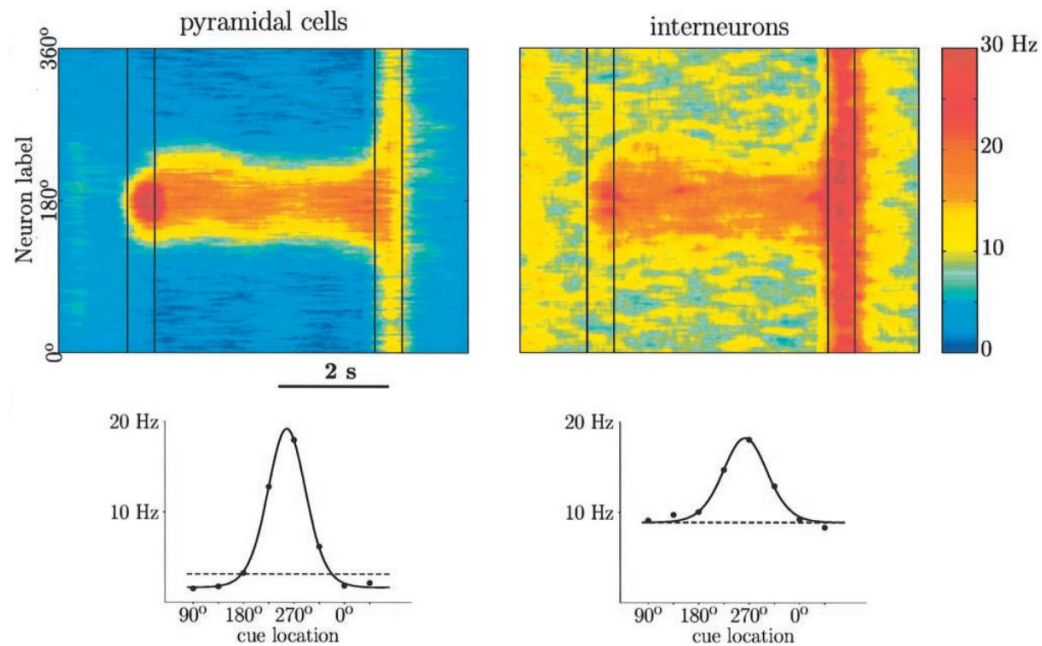


Figure 5. Bump attractor model of working memory in the prefrontal microcircuit. Heatmap shows individual neurons' (ordered along the y-axis by their preferred stimulus location) firing rates at each point of time. Vertical lines mark on- and offset of inputs, and the on- and offset of a strong, unspecific stimulus (marking the response). Before stimulus onset and after the response, neurons fire with spontaneous rates of ~4 Hz. During the delay, a bump of activity with a characteristic width and firing rates of ~20 Hz maintains stimulus-specific information. Background noise causes the bump to drift slightly over the course of the delay, a phenomenon related to delay-dependent imprecision (see Chapter 2). Bottom, tuning curves or memory fields for an example excitatory and inhibitory cell. Figure from Compte et al. (2000) with permission from Oxford University Press.

In the last decade, several studies have tested the predictions of attractor models of working memory in monkey PFC, finding evidence in line with both stable delay-firing, and the network origin of such activity (Wimmer et al. 2014; Inagaki et al. 2019; Kim et al. 2017). Yet, the fundamental assumptions of these models have been questioned recently (Lundqvist et al. 2018). Criticism is centered around the concept of persistent delay activity, claiming that it either reflects preparatory motor activity (which would be drastically reduced in delayed match-to-sample tasks (Shafi et al. 2007)), or a confound of trial-averaging of variable and transient single-trial activity. While the first concern poses a serious restriction to the interpretation of delay activity, the latter is answered by the compatibility of reverberatory, persistent activity with single-neuron, trial-to-trial variability and heterogenous temporal profiles: As outlined above, persistent activity likely is a network phenomenon, and does not need to occur in single cells. Accordingly, it has been demonstrated that while heterogenous on the single-neuron level, delay activity is best described by a stable code on the population level (Murray et al. 2017; Parthasarathy et al. 2019; Mendoza-Halliday and Martinez-Trujillo 2017).

How does persistent delay firing translate to human neuroimaging and EEG?

Interpreting human neuroimaging or human electroencephalography (EEG) studies in terms of underlying circuit mechanisms can be problematic, as the firing patterns of (delay-active) neurons are not unambiguously translated into neural activity measurable in EEG or fMRI BOLD responses. In fMRI, activity of a single voxel is best understood as the several-seconds-delayed, time-averaged, and spatially averaged activity of ca. 300,000 neurons that lie within several millimeters squared of cortical surface (Heeger and Ress 2002). For EEG, the relation to neural firing patterns is even more indirect: The potential measured at a scalp electrode reflects averaged spiking activity, but also subthreshold post-synaptic potentials from large, spatially distributed groups of neurons (Snyder and Smith 2015). It is non-trivial to reconstruct the source of a specific signal, due to physiological and mathematical difficulties: On the one hand, the directions along which EEG signals travel are very heterogeneous for different cortical regions, and signals distorted as they travel through inhomogeneous masses of neural tissue and scalp; on the other hand, there are more potential sources of the signal than electrodes or measurements, which makes the problem an underdetermined system (Grech et al. 2008). In any case, it has to be kept in mind that signals from different brain regions are integrated in potentials measured at a local electrode, and that EEG should therefore be understood on a whole-brain or global network level. Finally, neural computations can translate into temporally complex patterns in voltage traces (Snyder and Smith 2015), e.g. as a function of the synchrony in neural firing. However, the rules of this translation are not well known, further complicating inferences about neural processes from EEG signals.

While it is important to keep these limitations in mind, EEG offers an affordable and non-invasive way to study neural correlates of cognition in humans. Especially for the study of cognition in psychiatric disease that does not easily translate to animal models, it is often the only option to access brain activity in patients, together with fMRI. Moreover, EEG has a high temporal resolution and signals are relatively immediate (with a lag on the scale of tens of milliseconds), making it a suited instrument to study temporal aspects of neural responses (e.g., changing vs. stable codes, latency effects, etc.). While the EEG typically measures transient brain responses, other approaches have shown that sustained processes are reflected in the EEG signal, too: For example, when decomposing signals into oscillations of different frequencies, the power of the “alpha” band (8-12 Hz⁴) correlates

⁴In some studies, the alpha band refers to frequencies between 8-15 Hz.

with anticipatory or ongoing spatial attention (Banerjee et al. 2011; Foster and Awh 2019; Gould et al. 2011). Importantly, increases or decreases in the alpha band are retinotopically organized, so that their distribution across the scalp can be used to “decode” the spatial focus of attention. In fact, using this feature, alpha power has been recently employed to precisely decode spatial working memory contents over delays of several seconds (Foster et al. 2016). Similarly, but possibly during less extensive time periods, working memory contents can be decoded from multivariate raw voltage traces, where several properties of EEG memory codes correspond to memory codes measured from neural ensembles (e.g. Wolff et al. 2020).

What is the correspondence between these distributed, whole-brain memory codes and the circuit mechanisms described earlier? The sobering answer is, in most aspects, we do not know. Memory codes measured in alpha power possibly reflect the synchronization of neurons in location-tuned visual cortices (Kelly et al. 2006; Rihs et al. 2007), which could be driven by frontal feed-back inputs (Reinhart et al. 2012). When interpreting correlates of cognition measured in EEG in terms of circuit mechanisms, it is of advantage to compare findings to the non-human primate literature. Ideally, tasks should be designed in a way that make the comparison as straightforward as possible. Of course, these considerations are of lesser importance when interpreting neural correlates in EEG on their own account, outside the context of circuit mechanisms.

What is the role of synaptic plasticity in working memory?

In addition to the criticisms described in the last section, persistent activity holds other potential downsides: It is metabolically expensive (Attwell and Laughlin 2001), and relatively susceptible to distraction or interference from simultaneously held memories, as described in Chapter 1.2. Hence, it has been postulated that activity-dependent mechanisms with long time-scales other than increased firing rates, such as *short-term plasticity* (STP), might support attractor models to render memories more stable (Carter and Wang 2007; Itskov et al. 2011; Hansel and Mato 2013; Barbieri and Brunel 2008; Seeholzer et al. 2019; York and van Rossum 2009; Pereira and Wang 2015; Mongillo et al. 2012; Kilpatrick 2018; Yoon et al. 2020), and allow the temporary absence of persistent spiking during the delay (Mongillo et al. 2008; Fiebig and Lansner 2017).

These hybrid models are inspired by the identification of a range of STP mechanisms (Zucker and Regehr 2002), which span a spectrum of timescales from tens of milliseconds to several minutes, and can both *potentiate* or *depress* synapses. Evidence for STP in PFC has been found in vitro (Hirsch and Crepel 1990; Hempel et al. 2000; Wang et al. 2006) and in behaving animals (Fujisawa et al. 2008). In addition, specific features of STP in PFC (strong augmentation, strong facilitation in recurrent excitation) suggest that these mechanisms may play a role in the specific functions of PFC (Hempel et al. 2000; Wang et al. 2006). Typically, these studies considered STP as complementary to activity-based working-memory maintenance (Carter and Wang 2007; Itskov et al. 2011; Hansel and Mato 2013; Barbieri and Brunel 2008; Seeholzer et al. 2019; York and van Rossum 2009; Pereira and Wang 2015; Mongillo et al. 2012; Kilpatrick 2018; Yoon et al. 2020; Hempel et al. 2000; Wang et al. 2006), but the modeling work by Mongillo et al. (2008) has motivated a series of studies, most of them employing human neuroimaging techniques, that set out to demonstrate “activity-silent”, i.e. plasticity-based, maintenance of working memory by observing the absence of persistent activity during memory delays (e.g. Wolff et al. 2017; Rose et al. 2016). Apart from the issues with circuit interpretations of neuroimaging results exposed in the previous section, this line of argumentation is problematic from a science-theoretical perspective for two reasons: First, the absence of a mnemonic code in spiking activity of a specific area can only be observed if all neurons in that area were recorded, which to date is technically difficult with spike-recording techniques, and inaccessible to human neuroimaging techniques. Second, the absence of one mechanism does not per se prove the presence of a specific other mechanism, but rather the presence of any other mechanism (given the animal still successfully performs the task).

To conclude, it remains to be demonstrated that STP or other mechanisms with long-lasting activity-dependent alterations in neuronal excitability can replace persistent activity during working memory delays, and for the time being, stable memory codes as achieved by bistable neurons or attractor dynamics on the network level continue to be the most parsimonious explanation of how the brain maintains information in the absence of external inputs. At the same time, it is likely that some sort of long-timescale mechanisms coexist with firing-based neural codes, potentially stabilizing memory-related activity, reflecting early phases of memory consolidation in the neocortex, or merely reflecting residual phenomena of persistent firing. However, explicit evidence for this is still lacking.

1.2 Errors and systematic biases in working memory

As outlined in Chapter 1.1, working memory can be measured by how successfully an animal recalls the initially encoded stimulus. Depending on the characteristics of the task, different measures can be derived; the most popular ones being *capacity* (which implies the encoding of multiple items), *accuracy* (usually measured as percentage of correct responses) or *precision* (measuring how strongly responses vary around the stimulus value in continuous settings). Given the particular relevance of working memory tasks with continuous stimulus features, this section will focus on working memory errors as measured by imprecision (but generalize to errors measured by other metrics whenever it seems adequate). Errors in working memory can theoretically result from failed encoding, maintenance, or recall; therefore, only those phenomena that reflect increased memory failure for longer delay periods are considered maintenance-related (although different sensory features might expose memory decay more or less strongly (Pasternak and Greenlee 2005)). The systematic study of the sources of imprecision is interesting from a behavioral perspective itself. From a neuroscience perspective, it allows us to make inferences about the failure modes of neural mechanisms that hold working memory. In this sense, studying behavior is useful to test and adjust our mechanistic understanding of working memory. In Chapter 1.2, I will discuss the known sources of working memory imprecision and the neural basis thought to underlie the failure of working memory.

Forgetting

Working memory imprecision is explained by a multitude of factors, many of which are possibly still to be discovered. Imprecision can be measured by comparing encoded with reported stimulus values: Typically, most reports do reproduce stimuli relatively faithfully, but some reports are completely uncorrelated with encoded stimuli. These reports are thought to result from the most drastic of memory failures, forgetting, as a consequence of a breakdown of neural representations over the course of the delay (e.g. Funahashi et al.

1989). This phenomenon can be explained by insufficient excitatory reverberation, for example due to lower than optimal levels of recurrency between excitatory neurons (Cano-Colino et al. 2014). Another reason for forgetting could be the interference of strong new inputs with the currently maintained memory (Compte et al. 2000), causing memory-related neural activity to be abandoned (Sakai and Passingham 2003). However, the disruptive effect of distractors on persistent activity is stronger in parietal and inferotemporal regions than in PFC (Miller et al. 1996; Suzuki and Gottlieb 2013), so that distractors are typically successfully suppressed or filtered by PFC. In turn, if PFC persistent activity is abandoned, memory errors increase (Sakai and Passingham 2003). Finally, forgetting can be related to the number of stimuli that are memorized, resulting in an upper limit to working memory capacity (Zhang and Luck 2008; Miller 1956; Adam et al. 2017)⁵. In all cases, memories no longer retained by the responsible brain circuits will result in random responses at chance level, if the subject or animal is asked to report the stimulus after the delay (Zhang and Luck 2008).

Diffusion

A second source of errors in working memory is more subtle than the complete loss of memory-related activity, and does not result in forgetting: Neural circuits are subject to (intrinsic or extrinsic) noise that corrupts signals. In continuous attractor models such as the bump-attractor model proposed by Compte et al. (2000), the impact of noise on memory representations increases over time and contributes to a diffusion process of memory representations that is sometimes (somewhat non-orthodoxly) referred to as *random drift*⁶. The strength of this diffusion can be measured by the trial-to-trial variability of behavioral responses around the target, after eliminating random responses and systematic biases. As shown in Figure 6, increased memory delays lead to less precise responses (Funahashi et al. 1989; Rademaker et al. 2018). In both monkey and human electrophysiology, it has been shown that this phenomenon is consistent with randomly diffusing, but stable, neural codes (Wimmer et al. 2014; Wolff et al. 2020).

⁵ but see refs. (Gorgoraptis et al. 2011; Wilken and Ma 2004) for accounts of gradual corroboration and loss in precision, rather than forgetting, in multi-item working memory

⁶ The underlying assumption of a diffusion process is a continuous, so-called line attractor model of working memory. In a different class of attractor models, which are multistable with a discrete number of attractor states, remembered stimuli are categorized and stored as the typical class representation or fixed points (for a review, see Compte 2006). Due to the non-continuity of attractor states, diffusion or random drift should not affect memory representations in these models (Inagaki et al. 2019).

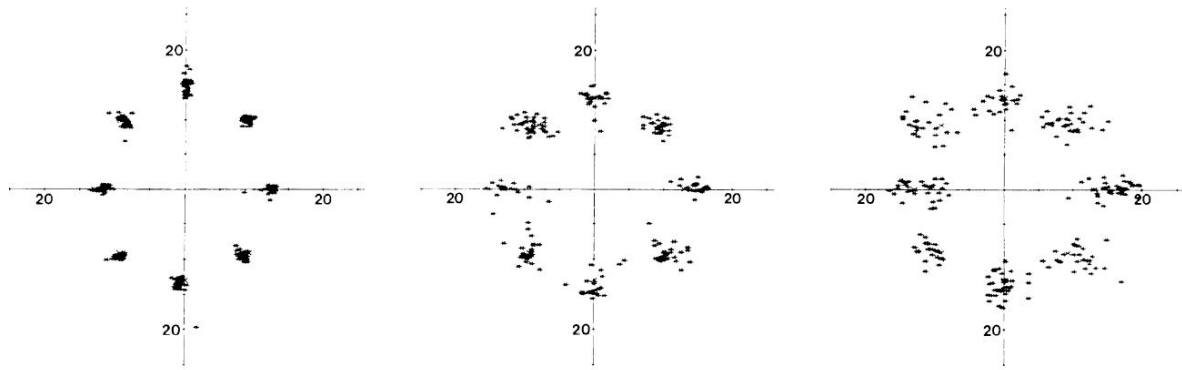


Figure 6. Saccadic endpoints after increasing delays in the experiment of Funahashi et al. (1989). The variability in monkeys' reports increases with increasing delay lengths, here observable in an increased spread of responses around the eight potential target locations. From left to right, visually guided saccades, saccades after 3 s delays, and after 6 s delays. Figure from Funahashi et al. (1989) with permission from The American Physiological Society.

Importantly, diffusion is not to be confused with (nonrandom) drift of memories towards specific stimulus values, a topic discussed in the next section. Both sources of error contribute to increased spread of reports around the target, usually resulting in a Gaussian-like distribution of reports relative to the target. Therefore, it is important to first explain errors that result from directed drift or different types of biases towards fixed stimulus values, before interpreting residual variability as the result of diffusion.

Systematic biases

The final part of this section has the goal to introduce the concept of biases in working memory, i.e. systematic distortions of memory representations towards or away from specific values of the parameter space that describes the stimulus. There is a large number of possible sources for bias, many of them probably still undiscovered, and I will only touch on a few of them: Bias towards fixed values of the parameter space, bias towards simultaneously presented stimuli, and bias towards previous experiences.

Again, all three of these classes are in agreement with attractor models of local working memory-maintaining circuits, such that memory representations drift towards specific values of the stimulus space. Some of these attracting values are fixed values which are constant over the course of an experiment (or even longer periods of time) and represent salient values of the stimulus distribution, such as visually salient colors in a color working memory task, or the cardinal directions in a visuospatial working memory task (Lipinski et al. 2010; Shin et al. 2017; Panichello et al. 2019; Bae et al. 2015). Attraction towards fixed

stimulus values shows a smooth profile that depends on the distance between the memorized stimulus and the value it is attracted to, and attractive effects typically increase with working memory delay (Shin et al. 2017; Panichello et al. 2019; Bae et al. 2015).

A second category of memory bias is attraction towards simultaneously presented stimuli, which can be observed in multi-item working memory (Almeida et al. 2015; Nassar et al. 2018), but also as an attraction towards distractors (Herwig et al. 2010; Rademaker et al. 2015; Van der Stigchel et al. 2007). Importantly, in both cases the attractive effect is again dependent on the relative distance between the remembered stimulus value and the stimulus value of the distractor, or the simultaneously remembered memory item, such that attraction is stronger for close-by distances. Almeida et al. (2015) have shown that attraction can be explained by (partial) merging of two bumps that represent different, but simultaneously held memories⁷.

Finally, biases can be directed towards previously experienced values. This attraction or repulsion can be a result of learning the stimulus statistics, so that the stimulus distribution serves as a Bayesian prior and causes an attraction towards the expected value of the distribution (Lieder et al. 2019; Ashourian and Loewenstein 2011). This bias is called contraction bias, as values are contracted towards the mean. Contraction bias is a specific example of the first class of biases described in this section, which are directed towards salient or prominent values of the stimulus space, and one could argue that the preference for certain stimulus values emerges from long-time learning of the sensory statistics in the outside world. While contraction biases could be explained in a local circuit model of working memory with (potentially learned) non-continuous attractor landscapes, similar to the model proposed by Panichello et al. (2019), Akrami et al. (2018) found that stimulus history in rats is stored separately from working memory delay activity in posterior parietal cortex (PPC).

⁷ Another phenomenon that occurs in multi-item settings are swap errors (Bays et al. 2009): Multi-item settings necessarily employ more than one stimulus dimension (such as color and location), so that at the time of recall, one feature can be used to indicate for which item the second feature has to be reported (“report the position of the red item (not the blue one)”). Swap errors occur when the non-probed item (in this example, the blue item) is reported instead of the probed one (the red one). This mix-up is thought to result from failures in feature-binding rather than memory bias, and has to be controlled for by calculating errors not only with respect to the target stimulus, but also to non-target stimuli, revealing an above-chance precision in case swap errors occurred.

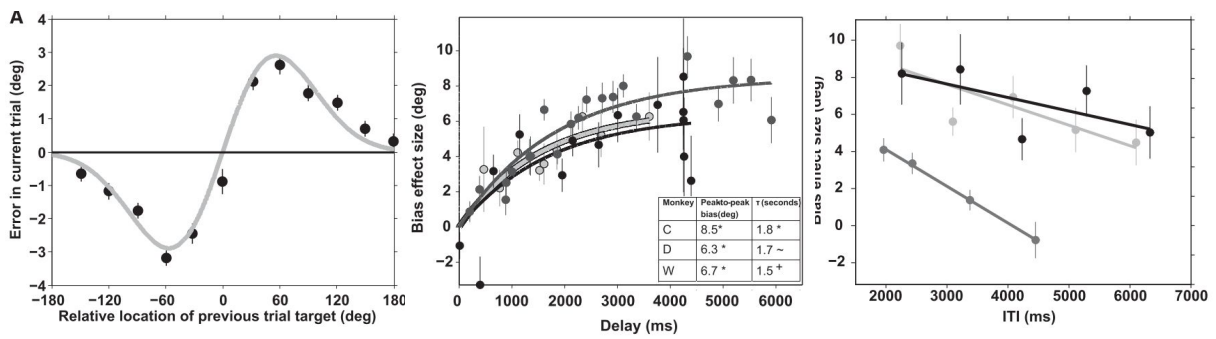


Figure 7. Left: Biases towards previous stimuli depend non-linearly on the distance between previous and current stimulus values. Middle: Serial dependence increases with increasing delay lengths. Right: For increased ITI durations between two trials, serial dependence decreases. Figure adapted from Papadimitriou et al. (2015) with permission from The American Physiological Society.

Biases towards previously experienced values do not necessarily imply optimal or Bayesian processing in the context of the working memory task: In tasks with a flat prior and uncorrelated trial-to-trial stimulus statistics, responses in human subjects, nonhuman primates, and rodents are attracted to the stimuli presented in the immediate past (Akrami et al. 2018; Papadimitriou et al. 2015; Hermoso-Mendizabal et al. 2020; Fischer and Whitney 2014). This type of biases is also called *serial dependence* in the human literature, as biases are short-lived and fade in the course of a few trials (Fischer and Whitney 2014), and they are weaker after longer inter-trial intervals (ITI) (Papadimitriou et al. 2015). There are other interesting temporal dynamics that characterize serial dependence: Attractive biases usually only emerge after working memory delays of several seconds, while biases are repulsive in the absence of memory requirements (Bliss et al. 2017). Repulsive biases are thus believed to be related to perception and sensory processing (Fritsche et al. 2017). Figure 7 shows the different temporal properties of serial dependence in a monkey behavioral study. These biases are small and systematic, and their different temporal features match with temporal characteristics of short-lived plasticity or cell-intrinsic, activity-dependent mechanisms with timescales of up to tens of seconds. Modeling work by Carter and Wang (2007) that includes such mechanisms in a continuous attractor model of working memory actually predicted the occurrence of biases towards previous stimuli, and theoretical work that explicitly tried to model the findings of Fischer and Whitney (2014) has successfully explained experimentally observed characteristics of serial dependence by including STP in continuous attractor models (Bliss and D’Esposito 2017; Kilpatrick 2018). The central idea in these models is that persistent firing facilitates or potentiates synapses

between delay-active neurons. After the response, in the absence of persistent activity⁸, these synapses remain potentiated for a few seconds, and thereby carry stimulus information from one trial to the next. When a new stimulus is presented, previously active neurons are more likely to fire, and memory-related activity is drawn to the respective stimulus values.

⁸ Note that Papadimitriou et al. (2017) have found that in a study with short ITIs, persistent firing continued until the next stimulus onset and attracted new memories through “active” interference, rather than through synaptic remnants.

1.3 The role of the NMDAR for working memory

In Chapters 1.1 and 1.2, I have introduced what is thought to be the neural basis of working memory maintenance, and what happens to cognition and behavior in case these mechanisms fail or are corrupted by noise or competing information. As pointed out in Chapter 1.2, failure and corruption of working memory results from a set of external factors, such as the length of the memory delay, the amount of concurrently presented stimuli, etc. More fundamentally however, a large number of internal variables of the neural circuit determines its suitedness and capacity for information maintenance. In Chapter 1.1, I touched upon variations of the cortical circuit architecture that can lead to persistent activity in PFC, but probably not in V1. Similarly, across individuals (or within individuals, as a result of long-term alterations), the same microcircuit can operate distinctly as a function of differences in *neurotransmitter* systems (Arnsten et al. 2012). Moreover, *neuromodulators* can flexibly move single neurons and circuits to a different operating regime (Marder 2012; Marder and Goaillard 2006), in which specific computations become possible or not⁹.

Pushing internal circuit parameters to specific directions in the parameter space can influence perception, cognition and behavior (Avery and Krichmar 2017), a fact exploited in pharmacological studies; and long-lasting alterations of neurotransmitter and neuromodulator systems are likely to be implicated in pathological brain function (Rolls et al. 2008; Arnsten et al. 2012; Cano-Colino et al. 2014). In this section, I will highlight the role of the NMDA receptor (NMDAR), which together with the AMPA receptor (AMPA) and the less frequent kainate and delta receptors is one of the main building blocks of the glutamatergic system (Dingledine et al. 1999). Glutamate is the main excitatory neurotransmitter, and its binding to these ionotropic receptors triggers their opening,

⁹ At the same time, neurons and neural circuits often show very similar behavior under widely distinct external and internal conditions. This observation led to the insight that if a neuron's behavior is subject to several parameters (e.g. if several different ion channels are expressed at a synapse), multiple solutions to the same neural behavior exist (Marder and Taylor 2011), a property that guarantees an inherent adaptability for certain parameter changes through compensation.

causing positively charged cations to enter the postsynaptic cell¹⁰. These positively charged currents eventually lead to the cell's depolarization and to an action potential or spike, if it depolarizes sufficiently. Changes in the glutamatergic system can affect circuit functions through reduced or increased synaptic excitation, which affects the so-called *excitation/inhibition balance* (E/I balance) of a system (discussed in the first part of this section). At the same time, the NMDAR plays a role in synaptic plasticity on different timescales, as discussed in the second part of this section. Finally, I will introduce two diseases that are linked to alterations in NMDAR density and function, anti-NMDAR encephalitis and schizophrenia, and discuss evidence for working memory dysfunction in relation to the NMDAR and to each of the two diseases.

The NMDAR contributes to persistent activity

While the NMDAR is best known for its features that support the induction of synaptic long-term potentiation (LTP) (Bliss and Collingridge 1993), it is also characterized by particularly slow deactivation kinetics that affect the decay time course of the excitatory postsynaptic current (EPSC) (Lester et al. 1990). In vitro NMDAR-mediated EPSCs are more prolonged in PFC than in V1 neurons, which might be explained by the predominant type of subunits that NMDARs in the respective cortex are composed of (with decay time constants τ : GluN2A < GluN2B \approx GluN2C \ll GluN2D; Cull-Candy et al. (2001)), and they are substantially more prolonged than AMPAR-mediated EPSCs in either area (Wang et al. 2008). As illustrated in Figure 8, the slow decay of EPSCs increases the postsynaptic neuron's excitability for a prolonged period, so that new inputs are integrated over time and increase the probability of repeated firing.

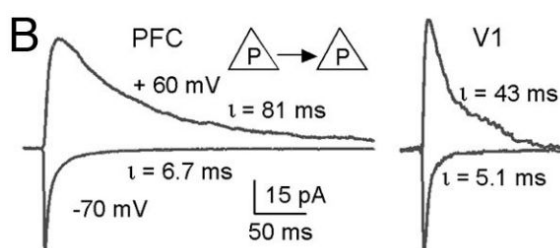


Figure 8. NMDAR- (upper trace) vs AMPAR-mediated (lower, inverted trace) EPSCs in PFC (left) vs. V1 (right). Due to their long deactivation time constant (here, $\tau = 81$ ms as compared to $\tau = 6.7$ ms for AMPARs), NMDARs in PFC favor the prolonged depolarization of postsynaptic cells, thereby increasing excitability and the probability of repeated firing. Figure from Wang et al. (2008). Copyright 2008 National Academy of Sciences, U.S.A.

¹⁰ With inward currents of sodium (Na^+) and calcium (Ca^{2+}) in case of the NMDAR, and (mostly) Na^+ in case of the AMPAR. The AMPAR is also characterized by an outward current of potassium ions (K^+).

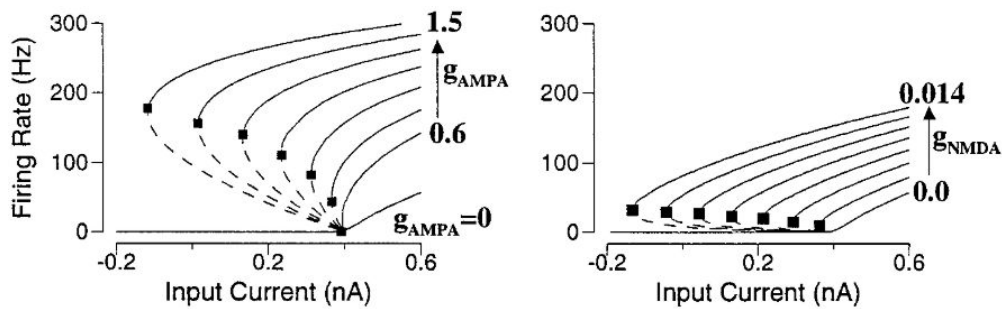


Figure 9. Frequency-current curves for different levels of AMPAR- (left) vs. NMDAR-mediated (right) synaptic coupling or conductance ($g_{\text{AMPA}}/g_{\text{NMDA}}$). The curves are described by $R = f(I_{\text{tot}}(R))$, where f is a function describing input-output relations, and $I_{\text{tot}}(R)$ the total input from recurrent and input synapses. Both AMPAR- and NMDAR-dominated networks are able to maintain stable states of spontaneous and persistent firing (solid lines). However, the transition to the persistent state (marked by filled squares), which is caused by a transient input current of ~ 0.4 nA, occurs at realistic firing rates (< 100 Hz) only in the presence of NMDAR currents. Moreover, sufficiently strong recurrent synaptic coupling of either the AMPA or NMDA type is necessary for solid bistability (space on the x-axis spanned by the dashed lines), as discussed in Chapter 1.1. Figure from Wang (1999), copyright 1999 Society for Neuroscience.

The PFC likely uses this property to generate persistent delay firing, a hypothesis formalized in spiking neural networks that explicitly model temporal kinetics of NMDAR and AMPAR at excitatory synapses (Wang 1999; Compte et al. 2000; Lisman et al. 1998). In these models, if the contribution of NMDARs is sufficiently strong, the EPSC of a single neuron is still above its baseline when inputs from recurrent connections arrive. Therefore, these models predict that repeated or persistent firing as a network phenomenon should crucially depend on NMDARs, given realistically low firing rates (< 100 Hz; Wang (1999)), as illustrated in Figure 9. In fact, recent work by Wang et al. (2013) confirmed this model prediction in an in vivo study in behaving monkeys: When selectively inhibiting NMDAR at excitatory synapses through a local application of NMDAR antagonist MK801, delay-active cells interrupted their persistent activity during the time of the pharmacological manipulation (Figure 10). Similarly, in the same study, the (less controlled but more broadly effective) systemic administration of MK801 and ketamine (another NMDAR antagonist) also affected delay firing, and ketamine reduced the monkey's precision in the delayed-response task. In line with these findings, several studies (Driesen et al. 2013; Honey et al. 2004; Anticevic et al. 2012) showed that the systemic administration of ketamine in humans reduced working memory-related changes in fMRI BOLD signals in PFC and affected working memory accuracy.

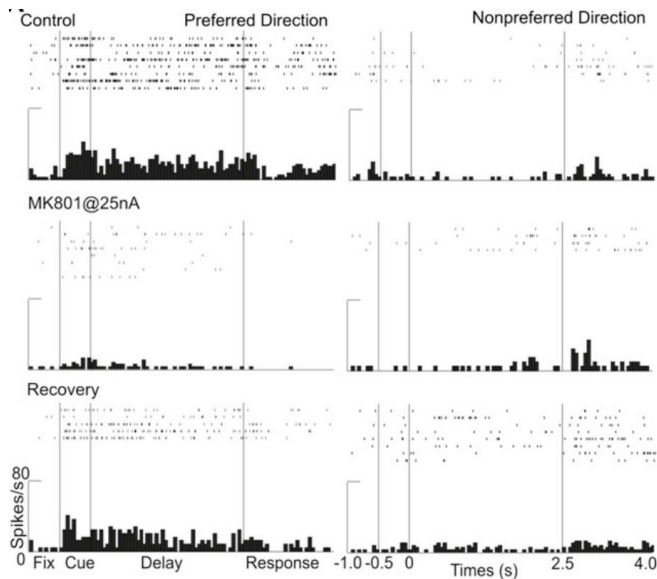


Figure 10. Raster plot and PSTH of delay-active cell in dIPFC of a monkey performing a visuospatial working memory task, before and after the local application (iontophoresis) of NMDAR antagonist MK801. In the control condition (no application of MK801) and after recovery from the drug manipulation, the delay cell increases its firing rate if the stimulus is shown in the preferred direction, and decreases its rate for the nonpreferred direction. Immediately after iontophoresis, the cell does not modulate its firing rate during the delay. Figure from Wang et al. (2013) with permission from Elsevier.

Synaptic plasticity on multiple timescales depends on the NMDAR

In addition to its *slow deactivation kinetics*, the NMDAR has a second prominent feature that makes it unique for the maintenance of memories (albeit on a different timescale than working memory). For the receptor to open, two conditions have to be met: The binding of glutamate and glycine, and a sufficiently advanced depolarization of the postsynaptic cell (Dingledine et al. 1999). The reason for this *voltage-gating* (in addition to glutamatergic ligand-gating) is the so-called magnesium (Mg^{2+})-block. A Mg^{2+} ion blocks the NMDAR and is only removed when the postsynaptic cell depolarizes (Nowak et al. 1984). Therefore, the NMDAR is sometimes referred to as a *Hebbian coincidence detector*: It activates only after the fast-opening AMPAR (or kainate receptor) already caused the depolarization of the postsynaptic cell and, coincidentally, additional glutamate is released upon presynaptic firing. Faithful to the Hebbian prediction, the synapses between co-active pre- and postsynaptic neurons are then strengthened locally, through a postsynaptic change in the number or efficacy of AMPAR that is triggered by the influx of Ca^{2+} through the open NMDAR (Malenka and Nicoll 1999). This NMDAR-dependent strengthening of synapses is called *long-term potentiation (LTP)* and is assumed to be the fundamental neural process involved in learning and long-term memory (Morris 1989).

Although most often studied in hippocampal (HC) synapses, it is widely accepted that LTP is a general property of excitatory synapses that takes place in many, and possibly all, neocortical areas (e.g., Castro-Alamancos et al. 1995; Kirkwood and Bear 1994; Crair and

Malenka 1995; Buonomano 1999; for reviews, see Feldman 2009; Mäki-Marttunen et al. 2020). LTP is studied through *in vitro* stimulation protocols, where the influence of repeated presynaptic stimulation on the EPSC of the postsynaptic cell is measured. After stimulation, EPSCs evoked by a probe stimulus can be higher (potentiation/LTP) or lower (depression/LTD) than EPSCs evoked before the stimulation protocol, depending on the exact experimental parameters (Dudek and Bear 1993; Malenka and Bear 2004). The effects of LTP can last hours to days, and probably contribute to the consolidation of memories (Clopath 2012; Clopath et al. 2008). In some protocols however, stimulation with insufficient duration, frequency or intensity to induce long-lasting LTP can cause more transient forms of potentiation that wane after a few minutes. Experiments that selectively blocked different subunits of the NMDAR in HC synapses showed that in fact, transient and long-lasting NMDAR-mediated potentiation might be partially triggered by different mechanisms: Volianskis et al. (2013) found that a quickly decaying form of transient LTP, also termed *short-term potentiation*, and long-lasting LTP both depended on receptors containing GluN2A/B subunits, while an intermediate form of short-term potentiation was disrupted by blocking receptors with GluN2B/D subunits. In light of these findings, an earlier study by Castro-Alamancos and Connors (1996) that applied the NMDAR antagonist AP5 to successfully disrupt short-term enhancement in somatosensory cortex could be interpreted as evidence for quickly decaying, GluN2A/B receptor-dependent short-term potentiation in neocortical and especially granular areas. Finally, Erickson et al. (2010) showed that NMDAR-dependent short-term potentiation (in HC) can be induced by as few as one or two presynaptic bursts, supporting the potential emergence of short-term potentiation during relatively short-lasting neural activity in flexible animal behavior, as compared to the long-lasting repeated stimulation in learning.

There is a range of STP mechanisms that can transiently enhance synaptic strength, many of which are not fully elucidated. As mentioned in Chapter 1.1, several processes at different time-scales have been described by Zucker and Regehr (2002), many of which appear to have an effect on the presynaptic probability of neurotransmitter release. In the context of working memory, the most popular mechanism is short-term facilitation, which results from residual calcium in the presynaptic terminal after neurotransmitter release (Tsodyks and Markram 1997). In contrast to the NMDAR-dependent short-term potentiation mechanisms described in the previous paragraph, this form of plasticity is more short-lived (up to 2 s), and more importantly, non-associative in the Hebbian sense. NMDAR-dependent, associative short-term potentiation mechanisms (Castro-Alamancos and

Connors 1996; Volianskis et al. 2013) also seem to increase the presynaptic probability of neurotransmitter release (but see Erickson et al. 2010), but probably either through a form of retrograde signaling¹¹ from the post- to the presynaptic cell (Volianskis et al. 2015; Meunier et al. 2017), or through presynaptic NMDARs (Corlew et al. 2008). Finally, their potentiating effects can last up to several minutes, and seem to depend on the frequency of probing, rather than the time since the potentiation protocol (Figure 11).

To conclude, synaptic plasticity on several timescales is mediated by NMDARs. Apart from the well-established findings for the implication of NMDAR-dependent LTP in learning and memory (Morris 1989), there is still little direct evidence for how the above-described NMDAR-dependent STP mechanisms are relevant in behaving animals. This is due to the difficulty of measuring plasticity in awake behaving animals. Moreover, and especially for STP, there is an added difficulty of disambiguating the influence of activity-based vs. plasticity-based contributions to animal behavior and cognition, due to the overlap of their timescales, and the overlap in underlying cellular mechanisms linked to the NMDAR.

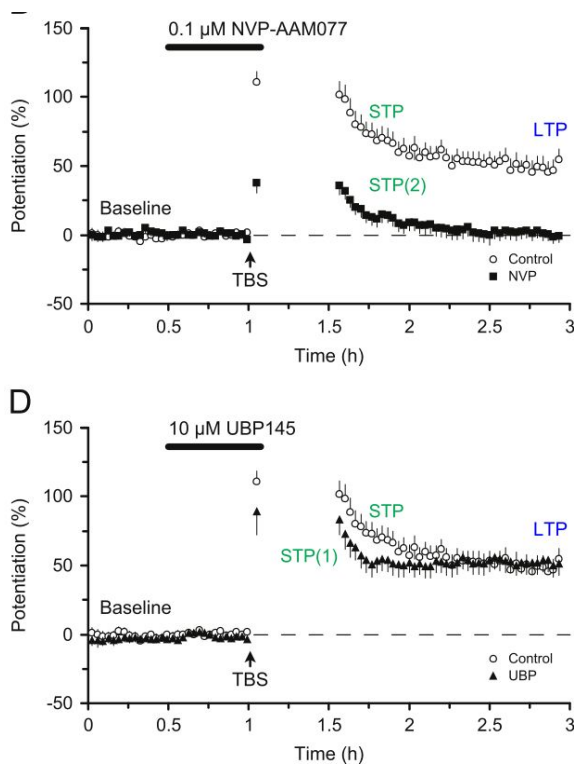


Figure 11. LTP and two forms of short-term potentiation (abbreviated in this figure as STP) depend on different NMDAR subtypes. When applying the NMDAR antagonist NVP (also AP5, not shown) that preferentially inhibits GluN2A/B containing receptors, LTP and the more short-lived form of short-term potentiation, STP(1), are disrupted (upper) at in vitro HC synapses. In contrast, antagonist UBP preferentially disrupts the longer-lasting form of short-term potentiation, STP(2), but not STP(1) or LTP (lower). The decay of STP(1) and STP(2) has time constants $\tau(1) = 7$ min and $\tau(2) = 16$ min at a stimulation rate of 0.067 Hz (Volianskis et al. 2013). However, the decay of short-term potentiation is activity-dependent, and in absence of probe stimulation, it can be maintained at a similar level for 30 min (or longer), as demonstrated by the 30 min break after theta-burst stimulation (TBS) that induced synaptic potentiation. Conversely, under higher frequencies as might occur in vivo, short-term potentiation will decay much faster (possibly on a seconds timescale). Figure from Volianskis et al. (2015) with permission.

¹¹ Potential candidates for retrograde signaling are neuromodulators nitric oxide (NO), brain-derived neurotrophic factor (BDNF; Meunier et al. 2017), or a retrograde flux of K^+ (Shih et al. 2013)

Schizophrenia

To conclude this section's review on how the NMDAR contributes to working memory, I will focus on two diseases that stand in the center of this thesis due to their link to NMDAR dysfunction: Schizophrenia and anti-NMDAR encephalitis. Specifically, after giving an overview of the clinical picture of each disease, I will highlight how (working) memory is affected in each of the two diseases, and how NMDAR dysfunction might contribute to the specific memory deficits.

Schizophrenia is a severe psychiatric disease that is mainly characterized by pervasive psychotic and cognitive symptoms. The onset of schizophrenia usually occurs in late adolescence or early adulthood, but genetic factors are a major predictor for the disease (Harrison and Weinberger 2005) and first-degree relatives share similar but milder symptoms in the cognitive domain (Sitskoorn et al. 2004). The most striking and therefore often first noted symptoms belong to the category of psychosis, and include hallucinations, delusions (false beliefs), disturbed thought patterns that lead to disorganized speech, and disorganized motor behavior. Cognitive symptoms include deficits in so-called "executive functions", such as planning, action inhibition, attention, and working memory (Barch and Ceaser 2012); but memory deficits on longer time scales also frequently occur, specifically in episodic memory (Pelletier et al. 2005).

Working memory deficits in schizophrenia are consistently found and might underlie more complex symptoms in cognition and behavior (Lee and Park 2005). Deficits occur across modalities (Forbes et al. 2009) and are related to maintenance, rather than the manipulation of memory contents (Barch and Ceaser 2012). The most robust findings include reduced working memory capacity (e.g., Gold et al. 2010; Hahn et al. 2018) and increased distractibility (Starc et al. 2017; Leonard et al. 2017; Gold et al. 2020) for people with schizophrenia, especially if distractors were similar to the working memory targets. Capacity and distractor effects in these studies appear to be independent of delay length, and might therefore not depend on gradual drift or diffusion of memory contents as described in Chapter 1.2. Similarly, Lee and Park (2005) and Gold et al. (2010) reported an absence of delay-dependent precision loss for patients with schizophrenia; however, Starc et al. (2017) and Gold et al. (2020) showed that given sufficiently long delays (~15-20 s), a loss in precision is indeed observed.

Some of these working memory deficits might be caused by an imbalance between cortical excitation and inhibition in people with schizophrenia, e.g. through a reduction in the activity of inhibitory interneurons (Lewis et al. 2005). The resulting disinhibition in a prefrontal circuit model, depicted in Figure 12, leads to increased memory drift, larger distractibility windows (Murray, Anticevic, et al. 2014), and potentially to a reduced capacity of the network to store multiple memory items. Disinhibition could follow from dysfunctional GABA receptors (GABAR; Lewis et al. 2005), which are the main cortical inhibitory receptors. Alternatively, the dysfunction of NMDAR at inhibitory interneurons could lead to cortical disinhibition (Anticevic et al. 2012; Cano-Colino and Compte 2012). In fact, there is steadily accumulating evidence for reduced NMDAR density and function in schizophrenia (including in PFC; Catts et al. 2016; Kristiansen et al. 2006), supporting the NMDAR hypofunction theory of schizophrenia that became popular in the last 20 years, after observations of schizophrenia-like symptoms and behavior in pharmacological studies with NMDAR antagonists (Olney et al. 1999).

Consistent with the NMDAR hypofunction model, working memory deficits and their neural correlates in healthy subjects under the administration of ketamine resemble those in subjects with schizophrenia (Driesen et al. 2008; Driesen et al. 2013). However, the attentive reader will notice that the blockade of cortical NMDAR at recurrent excitatory synapses, in contrast to the disinhibition model, could lead to a reduction in cortical excitation and a breakdown of memory-related delay activity (as explained earlier in this section; e.g., Wang et al. 2013). Depending on the synaptic site affected by the specific pharmacological manipulation, NMDAR blockade can lead to both disinhibition or deficient recurrent excitation in cortical circuits. Similarly, it is not clear in which direction cortical E/I balance in schizophrenia is perturbed, and different developmental stages of the disease could be linked to decreased or increased cortical excitation (Krystal et al. 2017).

Finally, by the NMDAR-mediated plasticity mechanisms described in the previous section of this chapter, NMDAR hypofunction could also affect LTP and short-term potentiation in schizophrenia. Evidence for this hypothesis comes from genetic studies in patients with schizophrenia that measure alterations in genes linked with synaptic plasticity (Harrison and Weinberger 2005), and conversely, from genetic animal models of schizophrenia that show working memory and long-term memory deficits *in vivo*, and reduced LTP and short-term potentiation in PFC and other neocortical areas *in vitro* (Arguello and Gogos 2012). Moreover, deficits in synaptic plasticity on the long term could

lead to alterations in circuit architecture, such as dendritic spine loss at recurrent synapses between layer 3 neurons (Glantz and Lewis 2000), and manifest in widely observed functional dysconnectivity in human neuroimaging studies (Stephan et al. 2006; Yang et al. 2016).

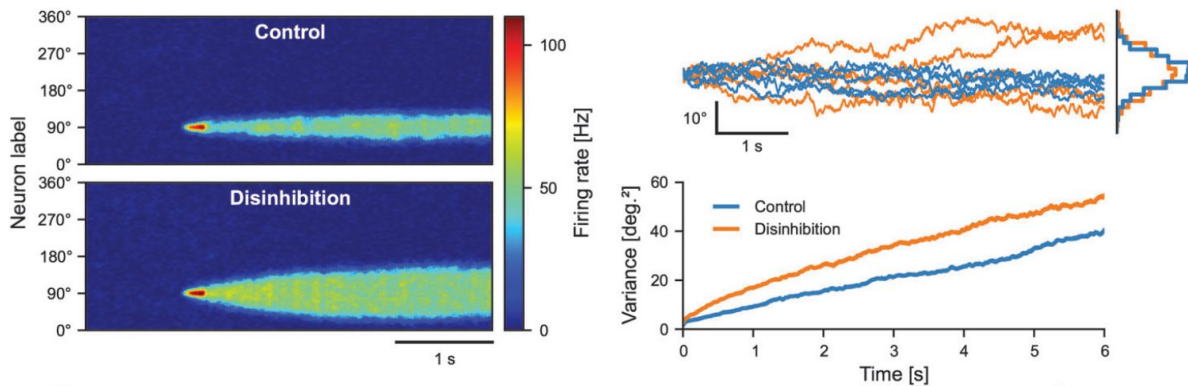


Figure 12. Decreased NMDAR-mediated conductance at inhibitory interneurons leads to disinhibition in a cortical circuit model of working memory. This disinhibition affects the persistent network state by broadening the bump of neurons that actively represent working memory content during the delay (left). Moreover, disinhibition causes an increase in noise in neural firing. This leads to stronger diffusion or random drift (upper right panel) over the course of the delay. Hence, over many trials, the distribution of responses is more variable for the disinhibited network than for the control condition, measured by the variance of network readout (center of mass of the bump), as shown in the lower right panel. Figure adapted from Murray, Anticevic, et al. (2014) with permission from Oxford University Press.

Anti-NMDAR encephalitis

Anti-NMDAR encephalitis is an autoimmune brain disorder that presents with neurological, psychotic and cognitive symptoms (Dalmau et al. 2007; Finke et al. 2012). The key mechanism of this disease is the formation of antibodies that target GluN1 subunits of the NMDAR and cause their internalization (Hughes et al. 2010; Dalmau et al. 2008), as depicted in Figure 13. Over the course of several weeks, the immune response affects an increasing number of NMDARs, causing a progressive deterioration of the clinical picture. Often, the disease starts with flu-like prodromal symptoms, followed by psychotic symptoms similar to those observed in schizophrenia, and cognitive deficits of working memory and speech. If left untreated, patients develop severe neurological symptoms, amongst others including abnormal or stereotyped movements, seizures, dysregulation of cardiac and respiratory function, and eventual coma. Compatible with the role of the NMDAR in LTP and memory formation, the most severe stages of the disease are accompanied by anterograde amnesia. Fortunately, the disease is easily diagnosed by the identification of antibodies in cerebrospinal fluid (CSF) or serum, and can be treated with

immunotherapy¹² (Dalmau et al. 2011). Complications can arise from early, frequently occurring misdiagnosis as schizophrenia or another psychotic disorder, leaving the patient untreated (Steiner et al. 2013). After treatment, more than 75% of patients substantially recover over the course of several months (but sometimes also periods longer than one year). The most long-lasting deficits concern working memory and other executive functions (Finke et al. 2012).

Most assessments of long-term and working memory deficits in (recovering) patients with anti-NMDAR encephalitis (reviewed in McKeon et al. 2018) have been performed with standardized neuropsychological instruments. One study by Finke et al. (2012) measured working memory deficits in a delayed match-to-sample task for color, location, or their association (as in earlier described multi-item tasks). The authors found that in some cases, (location and association) working memory was still impaired ~2 years after treatment, especially for long delays. Still, based on the current literature, it is difficult to establish which aspects of working memory fail in recovering anti-NMDAR encephalitis patients.

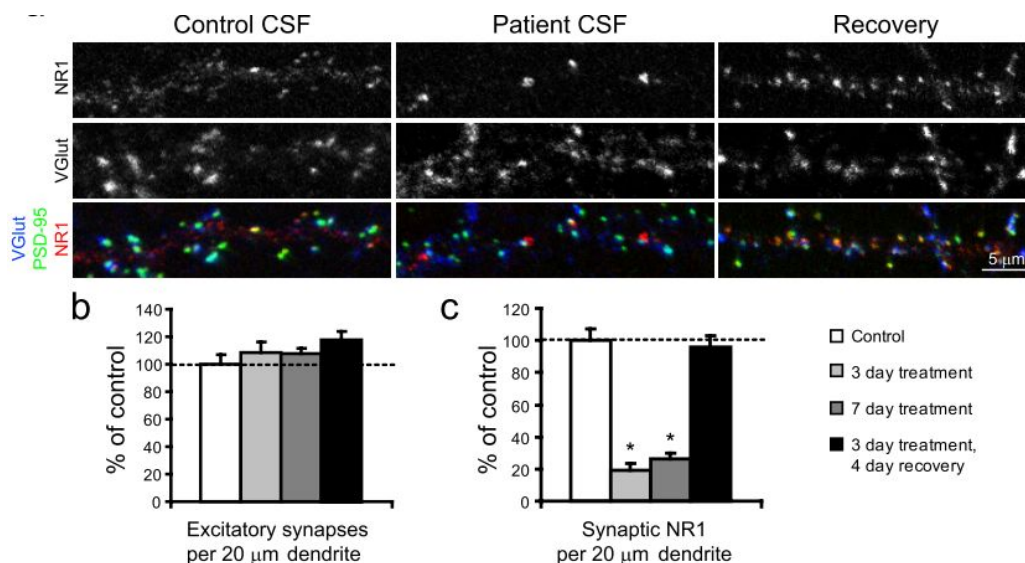


Figure 13. NMDAR density at rat HC neurons after the application of cerebrospinal fluid (CSF) from healthy control subjects (left), after application of anti-NMDAR encephalitis patients' CSF (middle), and after 3 days of patients' CSF followed by 4 days of recovery under application of control CSF. Receptor density was measured by immunostaining GluN1 subunits which are contained in all NMDAR (red dots). The percentage of GluN1 (NR1) clusters shows that receptor density is drastically reduced after the application of antibody-containing CSF, but recovers after the removal of patient-derived antibodies. In contrast, the number of excitatory synapses, measured by the colocalization of presynaptic protein VGlut and postsynaptic protein PSD-95, was not affected by applying patient CSF. Figure from Hughes et al. (2010).

¹² In ~50% of patients, the misdirected immune-response is triggered by an ovarian teratoma (tumor) that expresses NMDARs. In these cases, a surgical removal of the tumor should be performed together with immunotherapy (Dalmau et al. 2011).

Also, neural correlates of working memory failure in patients have not yet been assessed. However, the use of animal models that can be infused with patients' antibodies is a promising experimental approach both for the investigation of anti-NMDAR encephalitis, and for research on NMDAR function per se. When applying patients' antibodies to rat HC cells in vitro, a selective disruption of the NMDAR current can be observed in patch-clamp experiments (Hughes et al. 2010). This finding suggests that at least in *acute* stages of the disease, circuit mechanisms that depend on slow NMDAR-dependent currents, such as persistent activity (Wang 1999), should make a major contribution to working memory dysfunction. In contrast, long-term memory and its cellular substrate, LTP, show significant impairments in mice treated with antibodies from encephalitis patients: When patients' CSF was infused to ventricles of healthy mice (Planagumà et al. 2015; Planagumà et al. 2016), reversible memory deficits in a novel-object recognition task could be measured in vivo. When mice were sacrificed during the acute antibody treatment, a strong reduction in LTP was measured in stimulation protocols in vitro, as shown in Figure 14.

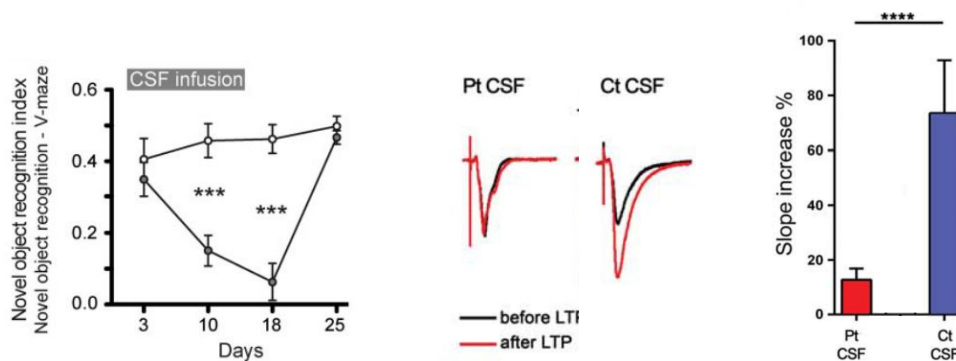


Figure 14. The infusion of human CSF containing NMDAR antibodies impairs long-term memory and LTP in mice. Left, the novel object recognition index measures the animal's long-term memory of objects. Higher indices indicate better memory. Mice were infused with patients' CSF for 14 days and showed progressive memory deficits. After the end of the treatment, memory performance recovered within a few days. Middle, field excitatory postsynaptic potentials (EPSPs) in mice infused with patients' or healthy controls' CSF, respectively, before and after an LTP stimulation protocol. While control mice showed substantially increased EPSP slopes (also, right panel), mice that were treated with patients' CSF did not. Compared to mice that were infused with healthy controls' CSF, significantly reduced LTP was induced in mice infused with NMDAR antibodies. Figure adapted from Planagumà et al. (2016) and Planagumà et al. (2015) with permission from John Wiley and Sons and Oxford University Press.

Chapter 2

Goals

The overarching goal of this thesis is to understand different sources of working memory errors as a function of underlying network computations in the “healthy” brain (Chapter 3.1), and in pathological conditions, namely anti-NMDAR encephalitis and schizophrenia (Chapters 3.2 and 3.3). I will especially focus on systematic memory biases that arise from the sequential structure of working memory tasks, as described in Chapter 1.2.

Specifically, Chapter 3.1 describes how persistent delay activity in working memory can lead to the formation of short-lived synaptic traces, and how persistent activity can be reactivated from these traces to influence upcoming memories. The section combines parallel data analyses in monkey prefrontal cortex and human EEG, and their connection to behavior. Moreover, we propose a computational model that captures prefrontal dynamics in monkey electrophysiology, and test predictions derived from this model in monkey and human data, and in a TMS perturbation experiment.

In Chapter 3.2, I describe behavioral working memory alterations in anti-NMDAR encephalitis and schizophrenia, as compared to healthy control subjects. I then interpret empirical findings in the framework of a prefrontal circuit model that combines network bistability through slow, NMDAR-mediated currents, and associative short-term potentiation that might depend on the NMDAR. The section scrutinizes whether working memory alterations observed in the two patient groups can be explained by a mere perturbation of E/I balance, or whether a dysfunction in short-term potentiation is needed to capture experimental results.

Chapter 3.3 explores neural correlates of behavioral working memory alterations in anti-NMDAR encephalitis and schizophrenia in an EEG study. In this section, will characterize patients' working memory codes during and between subsequent working memory trials, and test whether previous memories are reactivated in patients, similar to memory reactivations in healthy controls described in Chapter 3.1. Then, I will test whether differences in EEG memory codes reflect behavioral alterations in patients observed in Chapter 3.2, and relate the characteristics of the neural code back to the computational models proposed in Chapter 3.1 and Chapter 3.2.

Chapter 3

Results

The doctoral thesis of Heike Stein includes two published articles and one unpublished manuscript. In both of the two published articles, Heike Stein appears as first author, for her fundamental contributions detailed below. In addition, both articles have been published in top journals, in the highest decile of their scientific field. None of these articles has been used in another thesis, but the article in Nature Neuroscience is largely based on the material contained in the doctoral thesis of the other co-first author, Dr. Joao Barbosa. The specific contributions of Heike Stein and bibliometric details for each article are the following:

1) Joao Barbosa*, Heike Stein*, Rebecca L. Martinez, Adrià Galan-Gadea, Sihai Li, Josep Dalmau, Kirsten C. S. Adam, Josep Valls-Solé, Christos Constantinidis and Albert Compte. Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. (* equal contributions) *Nature Neuroscience* **23**:1016-1024, 2020 <https://doi.org/10.1038/s41593-020-0644-4>

Bibliometrics: WOS 2019 Journal Impact Factor 20.071, rank in Neurosciences 2/271

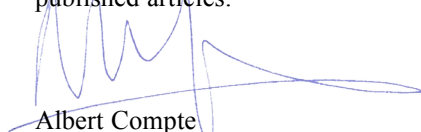
Author contributions: J.B. and A.C. performed the monkey data analyses. J.B. and A.C. developed the model. **H.S. and A.C. designed the human EEG research. H.S. and A.G.-G. performed the human EEG experiments. H.S., J.B. and A.C. performed the human data analyses.** A.C. and J.D. obtained the funding used for the human EEG research. K.C.S.A. performed the preliminary human EEG data analyses. J.B., R.L.M., J.V.-S. and A.C. designed the TMS experiments. R.L.M. performed the TMS experiments and performed the data analyses. S.L. performed the monkey experiments. C.C. designed the monkey research. **J.B., H.S. and A.C. discussed the results and wrote the manuscript.** All authors revised the manuscript and gave critical comments.

2) Heike Stein*, Joao Barbosa*, Mireia Rosa-Justicia, Laia Prades, Alba Morató, Adrià Galan-Gadea, Helena Ariño, Eugenia Martínez-Hernandez, Josefina Castro-Fornieles, Josep Dalmau and Albert Compte. Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia. (* equal contributions) *Nature Communications* **11**:4250,2020 <https://doi.org/10.1038/s41467-020-18033-3>

Bibliometrics: WOS 2019 Journal Impact Factor 12.121, rank in Multidisciplinary Sciences 6/71

Author contributions: **H.S., J.B., and A.C. designed behavioral and computational aspects of the study.** J.C.F., J.D., and A.C. designed clinical aspects of the study. **H.S. performed analyses of human behavior and computer simulations. H.S., J.B., and A.C. developed the computational model. H.S., A.M., L.P., and A.G.G. performed human experiments.** M.R.J. and L.P. performed neuropsychological testing. J.C.F., M.R.J., H.A., and E.M.H. recruited participants for the study. **H.S. and A.C. wrote the paper. H.S., J.B., J.D., and A.C. discussed the results and edited the paper.** All authors reviewed the paper for intellectual content.

All this satisfies the conditions contemplated in the program to present the doctoral thesis as a collection of published articles.



Albert Compte
October 2, 2020

3.1 Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory

In this section¹³, I will show that rather than operating independently, PFC persistent activity and “activity-silent” mechanisms interact dynamically to produce serial dependence in working memory, consistent with attractor models with synaptic plasticity.

Supplementary material for this section is included in Appendix A1.

¹³ This section has been published as:
Barbosa, J.*, Stein, H.*, Martinez, R.L., Galan-Gadea, A., Li, S., Dalmau, J., Adam, K.C.S., Valls-Solé, J., Constantinidis, C., & Compte, A. (2020). Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nat Neurosci* 23, 1016–1024. <https://doi.org/10.1038/s41593-020-0644-4>. (*equal contribution)



Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory

Joao Barbosa ^{1,8}, Heike Stein ^{1,8}, Rebecca L. Martinez¹, Adrià Galan-Gadea¹, Sihai Li², Josep Dalmau ^{1,3,4,5,6}, Kirsten C. S. Adam ⁷, Josep Valls-Solé¹, Christos Constantinidis² and Albert Compte ¹✉

Persistent neuronal spiking has long been considered the mechanism underlying working memory, but recent proposals argue for alternative ‘activity-silent’ substrates. Using monkey and human electrophysiology data, we show here that attractor dynamics that control neural spiking during mnemonic periods interact with activity-silent mechanisms in the prefrontal cortex (PFC). This interaction allows memory reactivations, which enhance serial biases in spatial working memory. Stimulus information was not decodable between trials, but remained present in activity-silent traces inferred from spiking synchrony in the PFC. Just before the new stimulus, this latent trace was reignited into activity that recapitulated the previous stimulus representation. Importantly, the reactivation strength correlated with the strength of serial biases in both monkeys and humans, as predicted by a computational model that integrates activity-based and activity-silent mechanisms. Finally, single-pulse transcranial magnetic stimulation applied to the human PFC between successive trials enhanced serial biases, thus demonstrating the causal role of prefrontal reactivations in determining working-memory behavior.

The mechanisms by which information is maintained in working memory are still not fully understood. Ample evidence supports a role for sustained neural activity in prefrontal^{1–3} and other cortices^{4,5}, possibly supported by attractor dynamics in recurrently connected circuits^{6,7}. However, recent studies have argued that memories may be maintained without persistent firing-rate tuning during memory periods⁸. This ‘activity-silent’ memory can be mediated by slowly decaying intrinsic or synaptic mechanisms, such as short-term synaptic plasticity^{9,10}, or by activity-dependent intrinsic mechanisms with a long time constant^{11–13} that could allow the reactivation of memories from latent storage. This computational proposal has received support from neuroimaging studies, whereby in some working memory tasks, despite good memory performance, stimulus information cannot be retrieved from neural delay activity, but later robustly reappears¹⁴ during comparison or response periods (but see also ref. ¹⁵).

The apparent incompatibility between activity-based and activity-silent memory maintenance has led to viewing them as exclusive alternatives⁸. However, modeling implementations of activity-silent conditions invariably require the network to be configured close to the same attractor regime⁹ that enables persistent activity. This attractor nonlinearity is necessary to increase the signal-to-noise ratio of the fading subthreshold signal for successful memory reactivation⁹. At the same time, activity-silent memory mechanisms may stabilize persistent activity in attractor networks (for examples, see refs. ^{11,16–18}). Interestingly, modeling studies have argued that the interaction of these mechanisms during the delay period would be reflected behaviorally in serial biases^{11,16}, but this theoretically appealing hypothesis still lacks experimental support.

Serial biases in spatial working memory denote small but systematic shifts of memory reports toward nearby locations memorized in the previous trial^{19–22}, which reveal a lingering representation of previous memories. Uncleared memory remnants have long been viewed as limiting working memory performance (proactive interference²³), but recent proposals suggest that they may be useful to inform working memory about the expected statistics in naturalistic conditions²⁴ (but see ²⁵), similar to other history biases with longer time scales and possibly different neural mechanisms (contraction bias^{26–28}). The functional relevance of biases implicates specific roles of higher-order brain areas. On the one hand, these areas could suppress maladaptive biases to minimize performance degradation^{29,30}. On the other hand, they might promote adaptive biases by maintaining a representation of stimulus history²⁶. Whether association areas generate or suppress serial biases in primates is currently undefined, and a mechanistic understanding of the generation of any type of history biases is still lacking.

Both attractor dynamics²⁰ and activity-silent^{11,16,31} mechanisms have been proposed to carry stimulus-selective information from one trial to the next to effect serial biases. However, dependencies of serial biases on inter-trial interval (ITI) durations^{30–22} are largely consistent with activity-silent and not activity-based mechanisms^{11,16,31}. Here, we sought to specify the interaction of activity-based and activity-silent PFC mechanisms in supporting serial biases while participants performed a spatial working memory task that engages attractor dynamics in the PFC⁶. Furthermore, this approach may offer indirect evidence that activity-silent and activity-based mechanisms co-occur during the delay period, as proposed by computational models (for examples, see refs. ^{11,16–18}).

¹Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. ²Department of Neurobiology and Anatomy, Wake Forest School of Medicine, Winston-Salem, NC, USA. ³Service of Neurology, Hospital Clínic, Barcelona, Spain. ⁴University of Barcelona, Barcelona, Spain. ⁵ICREA, Barcelona, Spain. ⁶Department of Neurology, University of Pennsylvania, Philadelphia, PA, USA. ⁷Department of Psychology and Institute for Neural Computation, University of California San Diego, La Jolla, CA, USA. ⁸These authors contributed equally: Joao Barbosa and Heike Stein. ✉e-mail: acompte@clinic.cat

Telling these mechanisms apart in the delay period is problematic because of their coactivation. By extending the relevant task periods to the ITI, we propose a way to disentangle them and to study the effect of their interaction on upcoming memories.

We compared the encoding properties of brain activity in the delay and ITI periods to identify the mechanistic basis of the memory trace that spans consecutive trials. We used behavioral and electrophysiological data collected in monkeys and humans: prefrontal multiple-unit recordings in monkeys and scalp electroencephalography (EEG) in humans. Between successive persistent activity mnemonic codes, we found an activity-silent code in the PFC that carried stimulus information through inter-trial periods. In addition, we found correlational and causal evidence, using transcranial magnetic stimulation (TMS), to indicate that fixation-period PFC reactivation from this activity-silent trace enhances attractive serial biases. These findings underscore the behavioral relevance of the dynamic interplay between attractor and subthreshold network dynamics in the PFC and reconcile these seemingly conflicting mechanisms. Our data suggests that this interplay could be the basis of closely associated memory storage processes operating at different time scales, thereby possibly serving different behavioral purposes.

Results

We trained four rhesus monkeys to perform an oculomotor delayed response task. The task consisted of remembering spatial locations at fixed eccentricity while maintaining fixation during a delay period of 3 s (Fig. 1a; Methods). The extinction of the fixation cue triggered the monkey to execute a saccade toward the remembered location and marked the beginning of a fixed ITI of 3.1 s, lasting until the appearance of the stimulus cue of the new trial (Fig. 1b). In addition, we tested 35 human participants in variations of the task performed by the monkeys (Methods). In all cases, we recorded the reported location and computed behavioral errors as angular distances to corresponding target locations. Following the methods described in previous studies¹⁹, we analyzed the dependence of the current-trial error on relative previous-trial location. Both monkeys and humans showed biased reports relative to previously remembered locations. These biases were attractive for short distances between previous-trial and current-trial locations, and repulsive for large previous–current distances (Figs. 1a and 2a). Our primary goal was to test the hypothesis that activity-silent and persistent activity working memory mechanisms interact to produce serial dependence effects. To this end, we investigated electrophysiological measurements in the ITI, including periods from the response to the subsequent fixation period.

Reactivation of previous memory information in the monkey dorsolateral PFC before new stimulus presentation. We collected single-unit responses from the dorsolateral PFC (dlPFC) of two monkeys while they performed the task. A substantial fraction of neurons in this area showed tuned persistent delay activity during the mnemonic delay period⁶ ($n = 206$ out of 822, Methods). These specific neurons are part of bump-attractor dynamics that characterize the memory periods of this task⁶. Based on this evidence, we assumed an attractor dynamics mechanism for persistent activity, and these terms are used interchangeably to refer to this network regime. Based on our hypothesis that an interplay of activity-silent and attractor mechanisms support serial biases, we focused our analyses on these neurons, and we grouped them in simultaneously recorded ensembles for decoding analyses ($n = 94$ ensembles, size range of 1–6 neurons; Extended Data Fig. 1a).

The firing rates of dlPFC neurons exhibited strong dynamics in the ITI compared to the characteristic stable dynamics during mnemonic delay periods (Fig. 1b). Phasic rate increases at response execution (R_{n-1} , Fig. 1b) and fixation onset (F_n , Fig. 1b) were hallmarks

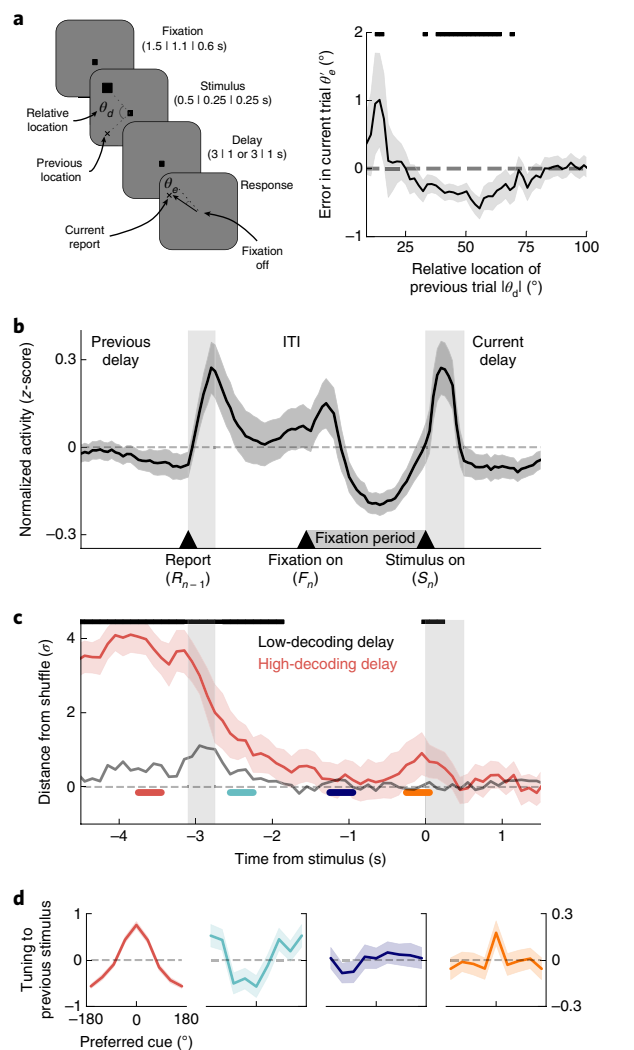


Fig. 1 | Previous-trial stimulus code reactivates before the forthcoming stimulus. **a**, General task design (left) and serial bias for four monkeys ($n = 11,670$ consecutive trial pairs; right). Trials with counter-clockwise previous reports relative to the current stimulus were collapsed into clockwise trials (folded errors, Methods). Positive (negative) values indicate response attraction (repulsion) toward previous locations presented at that relative distance from the current stimulus. Shading indicates bootstrapped \pm s.e.m. Black horizontal solid bars represent $P < 0.05$ (one-sided permutation test). Durations in different experiments are separated by vertical bars (monkey | EEG | TMS). **b**, Averaged, normalized firing rate of $n = 206$ neurons during the ITI (spike counts of 300-ms causal square kernel, z-scored in the interval $[-4.5$ s, 1.5 s]). Gray vertical bars mark the response and stimulus cue periods. **c**, The decoding accuracy of previous-trial stimulus from $n = 94$ independent ensembles, computed as the distance from the mean of the decoding accuracy in shuffled surrogates, in units of their standard deviation σ (Methods), averaged over ensembles with strong (red) and weak (gray) decoding in the delay period (Methods). Aligned with anticipatory ramping in late fixation (**b**), the previous-trial stimulus code reappears specifically in ensembles with stronger delay code (Extended Data Fig. 1). Black bars mark time points for which a decoding accuracy of 99.5% CI is above zero. **d**, Tuning to previous-trial stimuli, aligning responses to the preferred cue as defined in the delay period, and computed in different trial epochs (color-coded in **c**; two-sided bootstrap-test at preferred location: $P = 0.015$, $CI = [-0.3, -0.03]$, Cohen's $d = -0.17$ (cyan); $P = 0.865$, $CI = [-0.12, 0.14]$, Cohen's $d = 0.012$ (deep blue); $P = 0.025$, $CI = [0.024, 0.33]$, Cohen's $d = 0.15$ (orange); $n = 206$ neurons, shading depicts \pm s.e.m.). In all panels, unless stated otherwise, error shading marks bootstrapped 95% CI.

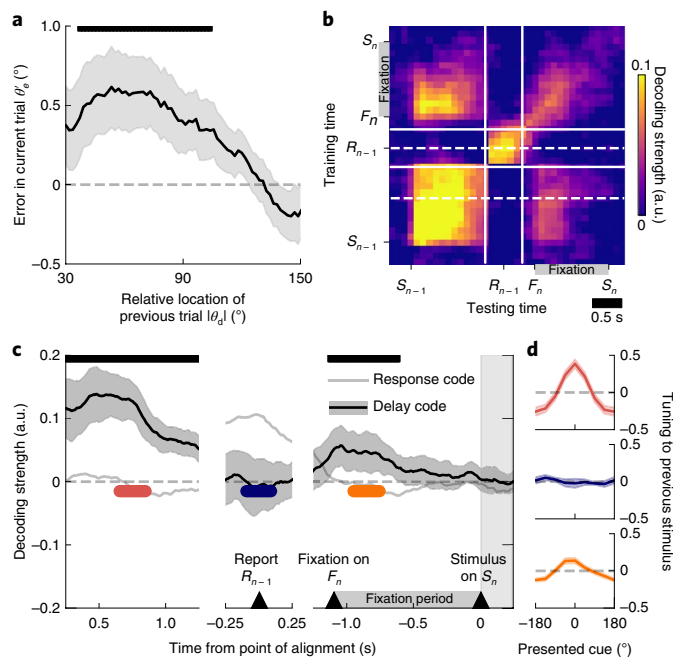


Fig. 2 | In human EEG, the delay code also reactivates in the fixation period.

a, Serial bias for human participants. Shading represents \pm s.e.m. **b**, Temporal generalization of previous-stimulus code for all combinations of training and testing times from previous-trial stimulus onset (S_{n-1}) and response (R_{n-1}) to current-trial fixation (F_n) and stimulus onset (S_n). Solid white lines mark the discontinuity of EEG fragments aligned to S_{n-1} , R_{n-1} and S_n . Dashed lines indicate the temporal center of transversal sections shown in **c**. a.u., arbitrary units. **c**, The decoding of previous stimulus during previous-trial delay (left), response (middle) and current-trial fixation period (right) for decoders trained during previous-trial delay (black line, 0.5 s–1. s after S_{n-1} , lower dashed line in **b**) and during previous-trial response (gray line, 0.5-s window centered on R_{n-1} , upper dashed line in **b**). The delay code is stable during the delay period, disappears during the response and reappears in current-trial fixation; see also **d**. In contrast, previous-trial response-related information is dynamic and not present in the fixation period. Error shading represents 95% CI. **d**, De-meaned reconstruction of tuning to the previous stimulus at different epochs for the delay decoder, marked in **c** (two-sided bootstrap-test preferred versus anti-preferred location: $P < 1 \times 10^{-6}$, CI = [0.55, 0.73], Cohen's $d = 3.6$ (red); $P = 0.69$, CI = [−0.22, 0.16], Cohen's $d = 0.10$ (blue); $P = 1 \times 10^{-6}$, CI = [0.17, 0.36], Cohen's $d = 1.35$ (orange); shading represents \pm s.e.m.). In **a** and **b**, the black horizontal bars indicate significant deviation from zero (bootstrap), $P < 0.05$ in **a**, $P < 0.005$ in **c** (both two-sided). For all panels, $n = 15$ independent participants.

in these dynamics, but we also noted an increase in the firing rate before stimulus presentation (S_n , Fig. 1b), which could reflect the anticipation of the upcoming stimulus due to fixed-length fixation periods. We wondered whether these rate changes were also related to dynamic changes in stimulus selectivity. Under the attractor-based hypothesis for serial biases³², sustained stimulus selectivity would be expected to extend from the delay period of the previous trial into the fixation period of the next trial. We measured selectivity by training a linear decoder on the spike counts of our neuronal ensembles and referenced its accuracy to that obtained by chance using a resampling approach (Methods). During the delay period, neuronal ensembles carried stimulus information and single neurons showed stimulus tuning (Fig. 1c,d, red). After report, the memorized location was still decodable from ensemble activity, but the tuning curves of single neurons showed a selective suppression of responses in their mnemonic preferred locations (Fig. 1c, cyan). This could reflect neuronal adaptation mechanisms or saccade

preparation toward the opposite direction to regain fixation. In the middle of the ITI, decoding accuracy was not different from chance and neurons were no longer tuned to the previous stimulus (Fig. 1c,d, deep blue), which suggests that the encoding of the previous stimulus had disappeared from neural activity. However, immediately before the presentation of the new stimulus and aligned with anticipatory ramping activity (Fig. 1b), the previous stimulus was again decoded and single-neuron tuning reappeared (Fig. 1c,d, orange). This reemergent stimulus information is consistent with previously-reported spiking selectivity during the ITI³², but we show here that there is a period in the ITI in which stimulus information cannot be decoded before it reappears at the end of the fixation period (late fixation). Furthermore, this code in late fixation is a reactivation of the representation active in the previous trial delay. This is supported by two pieces of evidence. First, information re-appearing occurred more strongly in those neuronal ensembles that maintained more stimulus information during the delay period (Fig. 1c; Extended Data Fig. 1). Second, the converging pattern of noise correlations at the end of the delay⁶ and in late fixation suggested a similar attractor-like network activation in both periods. Indeed, when the preceding stimulus appeared between the preferred locations of two neurons, these PFC neuron pairs exhibited negative noise correlations in late fixation (Extended Data Fig. 2). These negative noise correlations are a signature of a fixed-shape bump that diffuses from the initial stimulus location: as it moves closer to the preferred location of one neuron and away from the other, the firing rate increases for one neuron and decreases for the other⁶. Negative noise correlations appeared exclusively during late fixation, which strongly suggests that a bump is reactivated at that specific time point (Extended Data Fig. 2). Taken together, these results support that there is a reactivation of memory-period representation in the fixation period (reactivation period) following a period of absent selective neuronal firing in the dlPFC. This reactivation points at a relationship between mechanisms of delay memory encoding and mechanisms bridging the ITI to facilitate reactivation before the new stimulus.

Previous trial memory information reactivation in the fixation period of human EEG traces. In line with the monkey electrophysiology data, we found similar previous-trial traces in human EEG data ($n = 15$). We extracted alpha power from all electrodes and used a linear decoder to reconstruct the target location from EEG signals in each trial³³ (Methods). The target representation was significantly sustained during delay and response periods and in the fixation period of the next trial (Fig. 2b, diagonal axis). Importantly, at each time point, this dynamic EEG decoder uses signals originating from different cortical regions and could therefore combine temporally overlapping but spatially distinct representational components (for example, mnemonic versus response-related components). We therefore trained different linear decoders during the delay period (500–1,000 ms after stimulus onset, ‘delay code’) and around the time of the response (250 ms before to 250 ms after response, ‘response code’), and used the respective weights to extract previous-stimulus information throughout different periods of the trial (Fig. 2c). The delay code was stable during stimulus presentation and delay, but disappeared during the ITI, around the time of the response. In contrast, the response code did not generalize beyond the time at which the decoder was trained (Fig. 2c). We found that the delay code of the previous trial reappeared during the fixation period (Fig. 2c,d, orange), similarly to what we found in the monkey neurophysiology data (Fig. 1c), but slightly earlier in the ITI. In our human data, reactivation was possibly triggered by the onset of the fixation dot, while reactivation in the monkey PFC could be triggered by a ramping anticipatory signal in the fixed-duration ITI (Fig. 1b). These results provide a confirmatory correspondence with the time course of mnemonic decoding in the monkey data, but they also show the

temporal continuity between qualitatively distinct memory and response codes. The bidirectional transfer of information between memory and response representations in different brain areas could provide a bridge between the memory and reactivation periods observed in the PFC. Alternatively, response codes may just reflect the output motor commands, and mnemonic codes may subsist at a subthreshold level in the PFC to allow reactivations. We tested this hypothesis with a cross-correlation analysis of PFC units.

Increased cross-correlation suggests a latent trace during the ITI. We sought experimental validation for whether activity-silent mechanisms in the dlPFC still maintained stimulus information during the ITI between consecutive trials. We reasoned that if such latent activation (for example, a synaptic trace⁹) affected a group of interconnected neurons, these would be more likely to exceed their spiking threshold in synchrony^{8,34}. Following a preferred cue, neurons would increase their activity in the delay period and maintain latent activity-silent traces in the subsequent ITI that would be reflected in enhanced synchrony³⁴, but not enhanced rates. Moreover, we deduced that this reasoning was pertinent only to effective excitatory interactions (exc); that is, neurons interacting through effective inhibition (inh) should instead show a reduced probability of coactivation following a possible inhibitory efficacy enhancement by preferred stimuli in the previous trial³⁴.

To test this hypothesis, we selected pairs of neurons with similar selectivity ($n=67$ pairs, Methods) so that they had consistent activation (high or low firing rate) in the delay period. As per previous studies^{35,36}, we divided the selected pairs on the basis of their whole-trial cross-correlation peak sign in exc and inh interactions (Methods). We considered the following two conditions (Fig. 3a; Methods): trials in which the previous stimulus was shown close to either preferred location (pref; Methods) or far from preferred locations (anti-pref). Then, we computed a cross-correlation selectivity index (CCSI) by subtracting the amplitude of the central peak of the jitter-corrected cross-correlation function (coincident spikes within 20 ms; Methods, similar to ref. ³⁷) for pref and anti-pref trials for each neuron pair (Fig. 3b). Our hypothesis predicts positive (negative) CCSI for exc (inh) pairs in the ITI; that is, higher (lower) spike synchrony following preferred stimuli.

The CCSI computed in a period of the ITI where the firing rate had ceased to represent the stimulus (activity-silent period, Fig. 1c,d, deep blue) was positive, which reflects selectivity in neuronal synchrony to the previous stimulus for all interactions (Fig. 3c). We then investigated changes in CCSI values for exc and inh interactions across our two periods of interest: the activity-silent and reactivation periods (Fig. 1c, deep blue and orange, respectively). We found that their reactivation-period CCSI values significantly differed, being negative for inh interactions and positive for exc interactions (Fig. 3c). Finally, we explored the CCSI dynamics throughout the trial (Fig. 3d) and found that with the exception of immediately after the previous response, in which neurons showed anti-tuning to previous-trial stimulus (Fig. 1c), the CCSI for exc pairs was always positive, indicating stronger central-peak cross-correlation when the previous stimulus was preferred (Fig. 3d, orange). Conversely, for inh interactions, the CCSI was negative (stronger inh interactions following a preferred stimulus) only during reactivation and the previous-trial delay period (Fig. 3d, cyan), the periods in which PFC firing rates showed stimulus selectivity (Fig. 1c). This pattern is consistent with the latent memory mechanism residing in excitatory neurons and only being reflected in inhibitory interactions through collective engagement in bump-attractor dynamics during the delay period and at the time of reactivation. Importantly, this analysis was done during a period without firing-rate selectivity (Fig. 3f), thus free of a potential confound from firing rates (see Extended Data Fig. 3 for the same analysis performed during the delay period, where that caveat cannot be ignored.)

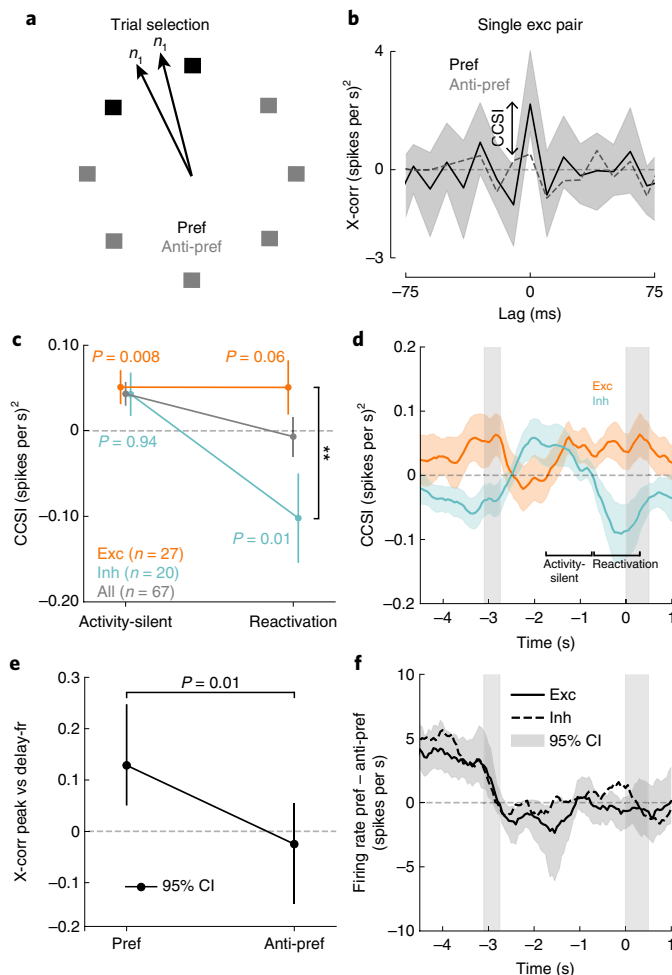


Fig. 3 | Cross-correlation selectivity to previous-trial stimulus suggests an activity-silent trace in the PFC.

a, Schematic of trial selection. For neuron pairs with a similar preferred location ($\leq 60^\circ$), we separated trials with stimulus near the preferred locations (pref) of the pair from trials with far locations (anti-pref). **b**, Cross-correlation (X-corr) of a sample PFC pair shows zero-lag peak selectivity to a previous-trial stimulus in the activity-silent period (one-sided permutation test, $P=0.025$, Cohen's $d=0.10$, $n=44$ independent trials). **c**, The CCSI was consistently positive in the activity-silent period, but became negative for inh interaction pairs during reactivation (two-sided permutation test, interaction period \times exc/inh, $P=0.03$, Cohen's $d=-0.6$). At reactivation, the CCSI for exc ($n=27$) and inh pairs ($n=20$) significantly differed (two-sided permutation test, $**P=0.006$, $d=0.75$). P values report results of one-tailed permutation tests according to our hypotheses (CCSI > 0 for exc, CCSI < 0 for inh). **d**, The CCSI in the ITI (1-s windows, 50-ms steps) for exc ($n=27$) and inh pairs ($n=20$). Except immediately after the report, where neurons show anti-tuning (Fig. 1d), the CCSI was positive for exc interactions. The CCSI was negative for inh interactions during previous delay and reactivation. Data were smoothed with a five-sample square filter. **e**, Trial-by-trial correlation between previous-delay spike counts for exc pairs and the ITI cross-correlation central peak (activity-silent period in **d**, Methods) is positive only for the pref condition (one-sided permutation test $P=0.017$, interaction $P=0.01$; $n=320$ and 769 trials for pref and anti-pref, respectively). **f**, The absence of a mean firing rate difference between the pref and anti-pref conditions (same pairs as in **d**) discards a confound between the rate selectivity and the CCSI. Error bars represent bootstrapped 95% CI (**b** and **e**) or s.e.m. (**c** and **d**).

This proves the existence of a latent trace of the stimulus in the PFC during the ITI, but it could still be reflecting selective subthreshold inputs from a different area that maintains tuned

persistent activity instead of selective local modulations in the PFC. To rule out this possibility and to strengthen the idea that stimulus information is directly transferred from an activity-based to an activity-silent code in the PFC, we tested whether the selectivity of exc interactions during the activity-silent period depended on the spiking activity of corresponding neurons in the previous delay period. Assuming a neuron-specific activity-dependent mechanism supporting the activity-silent code in the ITI, we predicted that the magnitude of the cross-correlation central peak in the activity-silent period would correlate on a trial-by-trial basis with the mean spike count recorded in the preceding delay period and specifically for pref (and not for anti-pref) trials (Methods). This prediction was confirmed in the experimental data (Fig. 3e). Thus, this cross-correlation analysis supports the hypothesis that previous, currently irrelevant, stimulus information remains in prefrontal circuits in latent states, undetected by linear decoders that do not take spike timing into consideration (Figs. 1c and 3f).

Bump reactivation as a mechanism for stimulus information reappearance. Based on our electrophysiology results and on prior modeling studies⁹, we formulated the bump-reattivation hypothesis to explain our data. We hypothesized that information held in memory as an activity bump during the delay period of the previous trial⁶ would be imprinted in neuronal synapses as a latent activity-silent trace during the ITI. This latent bump could be reactivated by the nonspecific anticipatory signal seen in the mean firing activity in the PFC (Fig. 1b) or by anticipatory mechanisms following an external cue that predicts stimulus presentation, such as the onset of a fixation dot (Fig. 2c). In fact, in a separate EEG experiment in which fixation lengths were jittered so as to make stimulus onsets unpredictable, we could not find any delay code reactivation (Extended Data Fig. 4).

To test the bump-reattivation hypothesis, we built a bump-attractor network model of spiking excitatory and inhibitory neurons. Based on our electrophysiology findings, short-term plasticity (STP) dynamics were included only in excitatory synapses (Methods). In each trial, stimulus information was maintained in activity bumps during the delay period by virtue of recurrent connectivity between neurons selective to the corresponding stimulus. During the ITI period, model neurons did not exhibit detectable tuning to the previous-trial stimulus (Fig. 4a, black, and Fig. 4b, deep blue)^{16,31}. However, the synapses of neurons that had participated in memory maintenance in the previous delay period were facilitated due to STP (Fig. 4a, deep blue). Parallel to our analysis presented in Fig. 3, this was reflected in the central peak of the ITI cross-correlation for pairs of excitatory model neurons, which maintained selectivity to the previous stimulus (Fig. 4a) even in the absence of single-neuron firing-rate selectivity (Fig. 4a, deep blue). We found that single-neuron tuning could be recovered from the hidden synaptic trace using a nonspecific input (drive) to the entire population (Fig. 4a,c; Methods, see also refs. ^{9,38}). Our biologically constrained computational model was therefore an explicit implementation of the bump-reattivation hypothesis that we had formulated.

The impact of bump reactivation on serial biases. We next used our computational model to derive behavioral and physiological predictions to test in our data, in particular in relation to serial biases. To simulate serial biases with our computational model, we ran pairs of consecutive trials with varying distance between the two stimuli presented in each simulation. We used the final location of the bump in the second trial (current-trial memory) as the ‘behavioral’ output of the model in that trial. We were able to model the profile of serial biases that were experimentally observed (Fig. 4d; Extended Data Fig. 5), similar to previous models^{16,31}. To test the impact of bump reactivation on serial biases, we compared

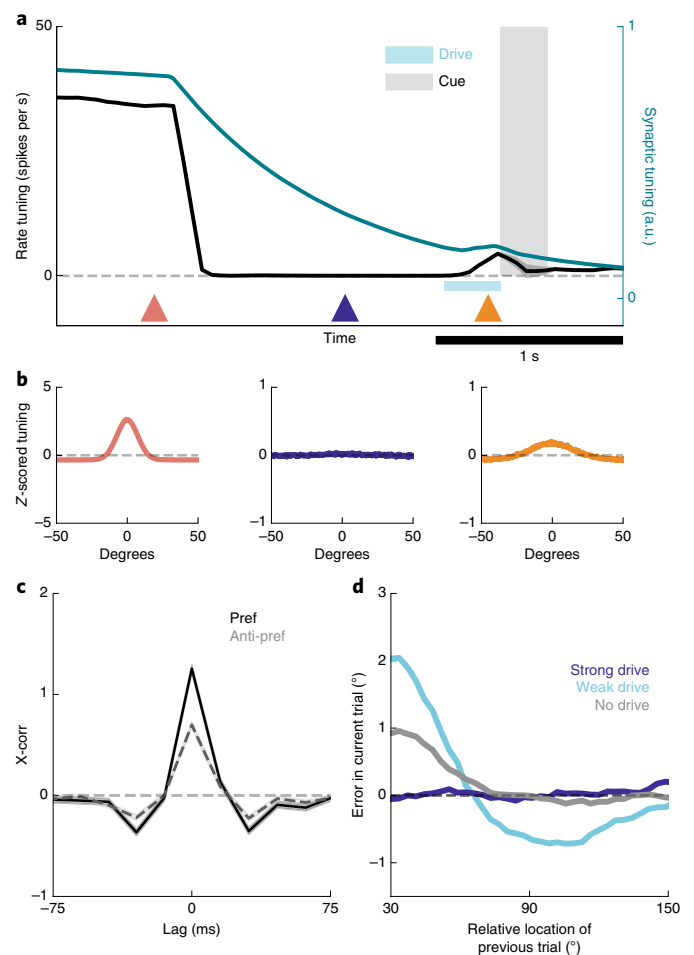


Fig. 4 | Bump-attractor model with STP accounts for serial dependence and neurophysiology. **a**, The average firing-rate tuning (black) and synaptic tuning (green) for 5,000 network simulations of two successive trials during the delay period (Methods). In the mnemonic period (red triangle), both rate and synaptic tuning are at their maximum, both driven by persistent bump-attractor activity (red plot in **b**). Following the memory period, a brief nonspecific hyperpolarizing input resets the baseline network state for the duration of the ITI (deep blue triangle and plot in **b**). This is reflected in a vanishing rate tuning, but long-lasting synaptic tuning that can regenerate firing-rate tuning (orange triangle and plot in **b**) through reactivation by a nonspecific input drive (cyan bar). **b**, Averaged single-neuron tuning to the previous-trial stimulus at different epochs, marked as colored triangles in **a**. **c**, Cross-correlation of model neurons in the ITI differed for the previous-trial stimulus in the preferred location (pref, black) and for anti-pref trials (gray) despite no firing-rate selectivity (**a** and **b**, deep blue). **d**, Serial bias plots computed from ‘behavioral response’ (Methods) in three different conditions of nonspecific depolarizing drive. A weak anticipatory drive increases attractive serial biases and produces repulsion from more distinct previous memories, while a strong drive removes serial biases.

the behavioral output of simulations with and without drive before the second trial stimulus (Methods). Bump reactivation resulted in stronger attractive biases for similar successive stimuli, and in repulsive biases for more dissimilar successive stimuli (Fig. 4d, cyan). We found that tuned intracortical inhibition^{39,40} was necessary for this emergence of repulsive biases after bump reactivation (Extended Data Fig. 5; see refs. ^{31,41} for an alternative mechanism). Finally, we tested the dependence of this behavioral effect on the strength of the nonspecific drive. A very short but strong impulse to the entire

network during the ITI quickly saturated all the synaptic facilitation variables, effectively removing all serial biases in the output of the network (Fig. 4d, deep blue). Thus, in this model, bump reactivation nonlinearly affects serial biases as the reactivation strength is varied. In summary, our model reproduced the behavioral and neurophysiological findings described in Figs. 1–3 and derived predictions concerning memory reactivations from silent traces that we then tested in the data.

Previous stimulus reactivation increases serial biases. The model predicts that higher reactivation of previous memories in the fixation period should be associated with stronger serial biases (Fig. 4d). We tested this prediction in our neural recordings from monkey PFC and in EEG recordings from the human scalp.

Monkey PFC. We first classified each trial on the basis of leave-one-out decoding of the previous stimulus trained and tested on activity from two different time windows during fixation: during a period with no stimulus information (activity-silent period; Fig. 1, deep blue) and at the time of reactivation (Fig. 1, orange). For each of these two windows, we separated high-decoding trials (first quartile) from low-decoding trials (all other trials) and computed bias curves separately. We found that serial biases were indistinguishable in the activity-silent period (Fig. 5a), but they were stronger for high-decoding than for low-decoding trials at the time of bump reactivation (Fig. 5b). This follows the prediction of our computational model, and it confirms the behavioral relevance of the bump reactivation before stimulus onset. This result was not dependent on a singular selection of trial separations, because for different proportions of high-decoding and low-decoding trials, the serial bias strengths (Methods) changed smoothly and remained consistent with the reported result (Extended Data Fig. 6). We then repeated the same analysis at different time points of the ITI. A significant difference in serial bias strength (Methods) emerged only when trials were classified as low-decoding versus high-decoding in the reactivation period (Figs. 1c and 5c, orange), and serial biases remained virtually indistinguishable at all other time points (Fig. 5c).

Human EEG. Analogous to the analysis of the monkey data, we grouped trials on the basis of their leave-one-out decoding accuracy of the previous stimulus (Methods). We separated high-decoding and low-decoding trials at two different time points: at the time of reactivation (Figs. 2 and 5f, orange) and at a fixation-period time point without stimulus information (activity-silent; Fig. 5c, black). Consistent with the monkey data and the prediction from our model, we found a stronger serial bias for high-decoding than for low-decoding trials for the reactivation period (Fig. 5e), but not for the activity-silent period (Fig. 5d), during which previous memory content was not decodable (Fig. 2c). The analysis was repeated for all other time points during the fixation period (Fig. 5f). Indeed, behavior exclusively depended on decoding accuracy at the time of delay code reactivation (Fig. 2, orange). Taken together, these results support the hypothesis that previous-trial memory reactivation before stimulus onset controls serial biases.

TMS-induced reactivations modulate serial biases. As a causal validation of the influence of pre-stimulus PFC reactivation on serial biases, we designed a TMS study. This is a relevant experiment because memory-dependent changes in human EEG alpha power cannot be unequivocally ascribed to a specific brain region, which limits the correspondence of our EEG and monkey dlPFC data. In particular, representations in larger and more organized occipital cortices might strongly contribute to visual EEG signals (for example, see ref. ³³), but could yet be driven by top-down projections from association cortices⁴². Inspired by a previous study¹⁴ that

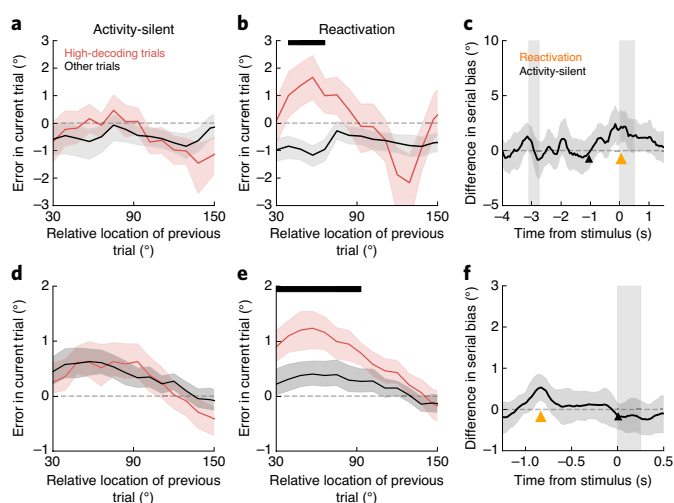


Fig. 5 | Bump reactivation from a hidden trace increases serial biases.

Serial bias for trials with high previous-trial stimulus information (upper quartile, red) and for all other trials (black) in monkeys (**a–c**, $n=1,362$ trials) and in humans (**d–f**, $n=15$ participants, with a range of 792–908 trials in this analysis). See Extended Data Fig. 6 for different quantiles. **a**, Trials selected based on a decoder trained and tested early in the fixation period (black triangle in **c**), did not reveal differences in serial bias. **b**, Serial biases were markedly enhanced for high-decoding trials when training and testing the decoder at the time of reactivation (Fig. 1c, orange triangle in **c**). **c**, Differences in serial bias curves between high-decoding and other trials became significant only in late fixation, concomitant with reactivation (Fig. 1c). Triangles mark the center of decoding windows for the splits shown in **a** and **b**. **d–f**, Same analyses for human EEG ($n=15$ independent participants). Note that for humans, **d** corresponds to an activity-silent period in late fixation (black triangle in **f**), and **e** to the reactivation period in early fixation (Fig. 2c, orange triangle in **f**). **f**, As for monkeys, serial bias differences in humans were significant only during reactivation. In **c** and **f**, time courses of differences between high-decoding and other trials were smoothed in time using a 5-sample (monkey) and 16-sample (human) square filter. Black horizontal bars (**b** and **c**) mark significant differences between high-decoding and other trials ($P < 0.05$, one-sided permutation test). Error shading represents 95% CI (**c** and **f**) or \pm s.e.m. (**a**, **b**, **d** and **e**).

reported reactivation of latent memories using TMS, we causally tested the role of the dlPFC in serial biases by applying single-pulse TMS during the fixation period. We had two control conditions to test our hypotheses: (1) we targeted the TMS coil at the dlPFC and the vertex in interleaved blocks, and (2) we randomly chose the TMS intensity in each trial (sham: 0%, weak-TMS: 70%, and strong-TMS: 130% of the resting motor threshold (RMT) of each participant; Methods). We found that TMS modulated serial biases when targeted at the dlPFC but not at the vertex (Fig. 6). Moreover, our computational model predicted a nonlinear dependence with stimulation strength (Fig. 4d), which was supported by the TMS data (Fig. 6b). Interestingly, the behavioral impact of PFC TMS stimulation declined throughout the session, as if participants became desensitized to the TMS pulse (Extended Data Fig. 7). Importantly, we show combined results from two separate experiments of $n=10$ participants each, one being a preregistered replication (Methods; Extended Data Figs. 8 and 9). These results provide causal evidence for the involvement of the PFC in the serial bias machinery during the ITI. Furthermore, we show that TMS affects serial biases in a nonlinear manner, as predicted by model simulations that implement the bump-rectivation hypothesis via the interplay of bump attractor and activity-silent mechanisms.

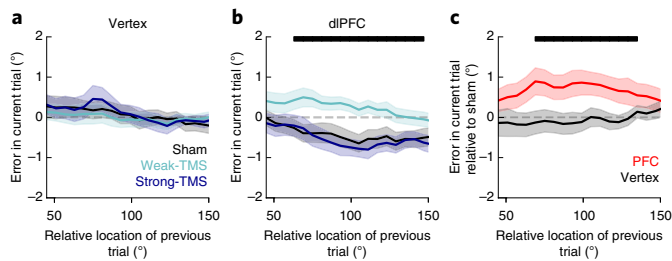


Fig. 6 | Single-pulse TMS on the dlPFC during fixation modulates serial biases nonlinearly. **a, b**, Serial bias computed in vertex (**a**) and PFC (**b**) blocks, separately for trials with strong fixation-applied TMS pulse (130% of RMT, deep blue), weak (70% RMT, cyan) and sham (0% RMT, black) for the first half of each session (225 trials, 2 sessions per participant, $n = 20$ participants; Extended Data Figs. 7–9). Serial biases were modulated by TMS in the PFC but not in the vertex (prev-curr \times TMS intensity \times coil location, $t_{18,272} = 2.21$, $P = 0.027$. For the dlPFC: prev-curr \times TMS intensity, $t_{11,087} = 2.13$, $P = 0.032$. For the vertex: $t_{7,166} = 0.03$, $P = 0.97$. Methods, linear mixed models; analysis performed on the entire session). In the PFC, serial bias modulation depended nonlinearly with the stimulation strength ($\Delta AIC = 4.6$, relative likelihood 0.9, for the comparison of regression models with nonlinear versus linear TMS intensity factor; Methods). **c**, The difference between serial biases computed for sham and weak-TMS trials in the vertex (black) and in the PFC (red) blocks. Error bars are bootstrapped \pm s.e.m. Solid black bars (**b** and **c**) mark significant differences (two-sided permutation test, $P < 0.05$, $n = 20$ independent participants).

Discussion

By studying the neural basis of serial biases, we showed how the interplay of bump-attractor dynamics and activity-silent mechanisms in the PFC maintains and eventually reactivates information about previous stimuli in spatial working memory. In delayed-response tasks, prefrontal tuned persistent activity consistent with bump-attractor dynamics characterizes the delay period and correlates with behavioral precision^{6,43}. We have now seen that this sustained activation disappears from the prefrontal network between trials, but is reactivated before the new trial (Figs. 1 and 2) and enhances behavioral serial biases (Figs. 5 and 6). This reactivation is directly linked to previous-trial activity: it emerged specifically in those neural ensembles that showed strongest persistent tuning in the delay period (Fig. 1c; Extended Data Fig. 1), it was decoded from human EEG data with decoders trained in the delay period (Fig. 2) and it exhibited the fingerprints of bump attractors as evaluated using pairwise correlations (Extended Data Fig. 2). Activity-silent mechanisms in the PFC bridge disconnected periods of persistent activity, carrying trial-specific information from one trial to the next (Fig. 3). Importantly, this latent tuning was directly associated with trial-by-trial firing rates in the preceding delay period (Fig. 3e), thus establishing a coupling between activity-based and activity-silent mechanisms in the PFC. Taken together, our results are consistent with the view that attractor-based and activity-silent mechanisms are jointly represented in the prefrontal circuit and that their tight interplay influences representations in spatial working memory. We specified this in a computational network model, whereby delay-period attractor dynamics imprint activity-silent mechanisms, which then retain information between trials and allow reactivations to recapitulate attractor states (Fig. 4).

Our data indicate that nonspecific PFC stimulation can revive subthreshold information, thus supporting the ideas put forward in computational models⁹ and in previous neuroimaging and EEG studies^{14,44,45}. Importantly, we obtained explicit causal evidence supporting the role of ITI reactivations in enhancing serial biases.

Similarly, recent causal evidence obtained in rodents²⁶ showed the role of parietal activations in generating history-dependent biases. However, the absence of selective mnemonic delay activity in rat parietal neurons²⁶ suggests that parietal ITI representations do not emerge from trace reactivations. A directed mechanistic investigation of the rat posterior parietal cortex in this task, similar to our efforts here, would be necessary to clarify the mechanisms and origin of history biases, and potential differences between the generation of contraction and serial biases in rodents and primates. More in line with our reasoning, human TMS studies found behavioral effects of memory reactivations when applied in the delay period, but only when memories were still behaviorally relevant¹⁴. In contrast, we show here that fixation-period TMS enhanced the behavioral influence of previous, already irrelevant memories. Reactivations may therefore not depend on behavioral relevance but rather on the decaying dynamics of activity-silent mechanisms; a more advanced decay of irrelevant memory traces may limit memory reactivations in ref. 14. Reactivations also offer alternative explanations to TMS effects in working memory that have previously been interpreted on the basis of network disruptions⁴⁶.

Our data support the idea that activity-silent and attractor-based mechanisms are not orthogonal, alternative mechanisms, but that they are interdependent mechanisms colocalized in the PFC. In turn, their different timescales may associate them preferentially with different types of memory processes. During active maintenance of working memory, rapid persistent attractor-based activity may encode memory, with slower activity-silent mechanisms providing a supporting, stabilizing role^{11,16,17}. Note that although direct evidence of this interplay in the delay period is problematic (Extended Data Fig. 3), our approach of separately assessing delay period and ITI, and their trial-by-trial correlation, indirectly supports this interplay and may be the most direct evidence that can be accessed extracellularly without resorting to detailed intracellular measurements in awake monkeys. After the deactivation of attractor-based active maintenance in the ITI, slowly decaying activity-silent maintenance may underlie secondary, possibly involuntary memory traces, leading to serial biases in upcoming trials. Note that previous studies have also proposed a central role for activity-silent maintenance for an additional, intermediate type of memory: unattended, behaviorally relevant memories^{14,44}. It was hypothesized that by resorting to different mechanisms, unattended memories may be reserved and protected while processing attended memories. Although our data do not address the mechanism of unattended memories, in our proposed framework, the close interplay between attractor-based and activity-silent mechanisms does not allow unattended memories (activity-silent memories) to be protected from intervening attended memories (attractor-based). This yields the prediction that serial-bias-like patterns of interference^{39,40} between unattended and attended memories should be observed in these experiments^{14,44}.

Our results have implications for the functional interpretation of serial biases and their relation with the interplay of prefrontal mnemonic mechanisms. First, enhanced serial biases after reactivating latent traces from earlier memories are consistent with the view that biases are the by-product of memory-supporting processes. As previous computational studies have shown, long-lasting cellular or synaptic mechanisms can enhance the stability of working memory retention (for examples, see refs. 11,16–18), but with the cost of across-trial interference of memories^{11,16}. Along these lines, a recently found reduction in serial biases in patients with schizophrenia⁴¹, anti-NMDA receptor encephalitis⁴¹ or autism²⁸ may reflect a reduced interplay of memory-supporting mechanisms. Second, we see an active role of the PFC in generating serial biases, rather than suppressing them as proposed by the proactive interference literature^{29,30}. This discrepancy could be resolved if the role of PFC was two-sided: (1) the PFC could generate biases either as a

by-product of stable memory retention^{11,16} or actively, in circumstances in which past memory traces are adaptive for behavior²⁴; alternatively, (2) strong PFC activation would suppress maladaptive memory remnants in situations where biases are particularly detrimental to behavioral performance. This dual PFC function is supported in our modeling and TMS data by the contrasting effect of weak and strong PFC activation on serial biases.

Our TMS experiment clarified our EEG results by demonstrating the role that the PFC plays in serial biases. Because we did not concurrently acquire EEG data during the TMS study, we could not directly measure the neural reactivation induced by the TMS pulse. However, prior work has shown the reactivation of EEG memory representations with TMS¹⁴, albeit in different conditions (pulses in the memory period targeted at parietal and occipital regions). Intriguingly, serial biases for trials without TMS stimulation in PFC-stimulation blocks were repulsive (Fig. 6b). We speculate that this was due to suppressive long-lasting physiological effects in the PFC that carried over from previous TMS-stimulated trials in the block¹⁷ (see Extended Data Fig. 10 for a phenomenological model of this hypothesis). Future work involving more fine-grained TMS intensities and carefully controlled block designs will be necessary to further clarify these results.

We proposed a computational model that can parsimoniously explain our data using STP in the synapses of a recurrent network. STP has also been used in previous computational models of interacting activity-based and activity-silent dynamics^{9,10,13} and of serial biases^{16,31}. Beyond previous modeling efforts, we explored the mechanistic requirements of code reactivations before a new trial, and we derived predictions whose validation conferred plausibility to the model. Our findings do not unequivocally identify this mechanism and we could have chosen another mechanism with a long time constant to computationally implement our hypothesis (for example, calcium-activated depolarizing currents¹⁷, depolarization-induced suppression of inhibition¹¹ or short-term potentiation⁴⁸). Also, synaptic plasticity mechanisms linked to feed-forward connections into the PFC³⁸ could conceivably play a role. Still, several lines of evidence support the involvement of STP in prefrontal function. First, there is explicit evidence for enhanced short-term facilitation and augmentation among PFC neurons in *in vitro* studies^{49,50}. Second, extracellular recordings in behaving animals cannot directly probe activity-silent mechanisms, but indirect evidence for synaptic plasticity has been gathered from prefrontal activity correlations of rodents engaged in working memory tasks³⁵. Our study also follows this approach to seek evidence for activity-silent stimulus encoding, but we applied it specifically at time periods without firing-rate codes for task stimuli, thus unambiguously decoupling activity-silent from activity-based selectivity (Fig. 3; Extended Data Fig. 3).

In summary, our data show that subthreshold traces of recent memories remain imprinted in PFC circuits and bias behavioral output in working memory in particular through network reactivations of recent experiences. Our findings suggest that the dynamic interplay between attractor and subthreshold network dynamics in the PFC supports closely associated memory storage processes: from effortful memory to occasional reactivation of fading experiences.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41593-020-0644-4>.

Received: 22 August 2019; Accepted: 21 April 2020;
Published online: 22 June 2020

References

- Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* **61**, 331–349 (1989).
- Kubota, K. & Niki, H. Prefrontal cortical unit activity and delayed alternation performance in monkeys. *J. Neurophysiol.* **34**, 337–347 (1971).
- Fuster, J. M. & Alexander, G. E. Neuron activity related to short-term memory. *Science* **173**, 652–654 (1971).
- Leavitt, M. L., Mendoza-Halliday, D. & Martinez-Trujillo, J. C. Sustained activity encoding working memories: not fully distributed. *Trends Neurosci.* **40**, 328–346 (2017).
- Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R. & Haynes, J.-D. The distributed nature of working memory. *Trends Cogn. Sci.* **21**, 111–124 (2017).
- Wimmer, K., Nykamp, D. Q., Constantinidis, C. & Compte, A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neurosci.* **17**, 431–439 (2014).
- Inagaki, H. K., Fontolan, L., Romani, S. & Svoboda, K. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature* **566**, 212–217 (2019).
- Stokes, M. G. “Activity-silent” working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* **19**, 394–405 (2015).
- Mongillo, G., Barak, O. & Tsodyks, M. Synaptic theory of working memory. *Science* **319**, 1543–1546 (2008).
- Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J. & Freedman, D. J. Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nat. Neurosci.* **22**, 1159–1167 (2019).
- Carter, E. & Wang, X.-J. Cannabinoid-mediated disinhibition and working memory: dynamical interplay of multiple feedback mechanisms in a continuous attractor model of prefrontal cortex. *Cereb. Cortex* **17**, i16–i26 (2007).
- Fiebig, F. & Lansner, A. A spiking working memory model based on Hebbian short-term potentiation. *J. Neurosci.* **37**, 83–96 (2017).
- Orhan, A. E. & Ma, W. J. A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nat. Neurosci.* **22**, 275–283 (2019).
- Rose, N. S. et al. Reactivation of latent working memories with transcranial magnetic stimulation. *Science* **354**, 1136–1139 (2016).
- Christophel, T. B., Iamshchinina, P., Yan, C., Allefeld, C. & Haynes, J.-D. Cortical specialization for attended versus unattended working memory. *Nat. Neurosci.* **21**, 494–496 (2018).
- Kilpatrick, Z. P. Synaptic mechanisms of interference in working memory. *Sci. Rep.* **8**, 7879 (2018).
- Tegnér, J., Compte, A. & Wang, X.-J. The dynamical stability of reverberatory neural circuits. *Biol. Cyber.* **87**, 471–481 (2002).
- Seeholzer, A., Deger, M. & Gerstner, W. Stability of working memory in continuous attractor networks under the control of short-term plasticity. *PLoS Comput. Biol.* **15**, e1006928 (2019).
- Fischer, J. & Whitney, D. Serial dependence in visual perception. *Nat. Neurosci.* **17**, 738–743 (2014).
- Papadimitriou, C., Ferdoash, A. & Snyder, L. H. Ghosts in the machine: memory interference from the previous trial. *J. Neurophysiol.* **113**, 567–577 (2015).
- Fritsche, M., Mostert, P. & de Lange, F. P. Opposite effects of recent history on perception and decision. *Curr. Biol.* **27**, 590–595 (2017).
- Bliss, D. P., Sun, J. J. & D'Esposito, M. Serial dependence is absent at the time of perception but increases in visual working memory. *Sci. Rep.* **7**, 14739 (2017).
- Jonides, J. & Nee, D. E. Brain mechanisms of proactive interference in working memory. *Neuroscience* **139**, 181–193 (2006).
- Kiyonaga, A., Scimeca, J. M., Bliss, D. P. & Whitney, D. Serial dependence across perception, attention, and memory. *Trends Cogn. Sci.* **21**, 493–497 (2017).
- Barbosa, J. & Compte, A. Build-up of serial dependence in color working memory. Preprint at <https://www.biorxiv.org/content/10.1101/503185v1> (2018).
- Akrami, A., Kopec, C. D., Diamond, M. E. & Brody, C. D. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature* **554**, 368–372 (2018).
- Hermoso-Mendizabal, A. et al. Response outcomes gate the impact of expectations on perceptual decisions. *Nat. Commun.* **11**, 1057 (2020).
- Lieder, I. et al. Perceptual bias reveals slow-updating in autism and fast-forgetting in dyslexia. *Nat. Neurosci.* **22**, 256–264 (2019).
- D'Esposito, M., Postle, B. R., Jonides, J. & Smith, E. E. The neural substrate and temporal dynamics of interference effects in working memory as revealed by event-related functional MRI. *Proc. Natl Acad. Sci. USA* **96**, 7514–7519 (1999).
- Feredoes, E., Tononi, G. & Postle, B. R. Direct evidence for a prefrontal contribution to the control of proactive interference in verbal working memory. *Proc. Natl Acad. Sci. USA* **103**, 19530–19534 (2006).
- Bliss, D. P. & D'Esposito, M. Synaptic augmentation in a cortical circuit model reproduces serial dependence in visual working memory. *PLoS ONE* **12**, e0188927 (2017).

32. Papadimitriou, C., White, R. L. & Snyder, L. H. Ghosts in the machine II: neural correlates of memory interference from the previous trial. *Cereb. Cortex* **27**, 2513–2527 (2017).
33. Foster, J. J., Sutterer, D. W., Serences, J. T., Vogel, E. K. & Awh, E. The topography of alpha-band activity tracks the content of spatial working memory. *J. Neurophysiol.* **115**, 168–177 (2016).
34. Trousdale, J., Hu, Y., Shea-Brown, E. & Josić, K. Impact of network structure and cellular response on spike time correlations. *PLoS Comput. Biol.* **8**, e1002408 (2012).
35. Fujisawa, S., Amarasingham, A., Harrison, M. T. & Buzsáki, G. Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nat. Neurosci.* **11**, 823–833 (2008).
36. Barthó, P. et al. Characterization of neocortical principal cells and interneurons by network interactions and extracellular features. *J. Neurophysiol.* **92**, 600–608 (2004).
37. Cohen, J. Y. et al. Cooperation and competition among frontal eye field neurons during visual target selection. *J. Neurosci.* **30**, 3227–3238 (2010).
38. Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T. P. & Husain, M. Neural mechanisms of attending to items in working memory. *Neurosci. Biobehav. Rev.* **101**, 1–12 (2019).
39. Almeida, R., Barbosa, J. & Compte, A. Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. *J. Neurophysiol.* **114**, 1806–1818 (2015).
40. Nassar, M. R., Helmers, J. C. & Frank, M. J. Chunking as a rational strategy for lossy data compression in visual working memory. *Psychol. Rev.* **125**, 486–511 (2018).
41. Stein, H. et al. Disrupted serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia. Preprint at <https://www.biorxiv.org/content/10.1101/830471v1> (2019).
42. Reinhart, R. M. G. et al. Homologous mechanisms of visuospatial working memory maintenance in macaque and human: properties and sources. *J. Neurosci.* **32**, 7711–7722 (2012).
43. Sajad, A., Sadeh, M., Yan, X., Wang, H. & Crawford, J. D. Transition from target to gaze coding in primate frontal eye field during memory delay and memory-motor transformation. *eNeuro* **3**, ENEURO.0040-16.2016 (2016).
44. Wolff, M. J., Jochim, J., Akyürek, E. G. & Stokes, M. G. Dynamic hidden states underlying working-memory-guided behavior. *Nat. Neurosci.* **20**, 864–871 (2017).
45. Bae, G.-Y. & Luck, S. J. Reactivation of previous experiences in a working memory task. *Psychol. Sci.* **30**, 587–595 (2019).
46. Zokaei, N., Manohar, S., Husain, M. & Feredoes, E. Causal evidence for a privileged working memory state in early visual cortex. *J. Neurosci.* **34**, 158–162 (2014).
47. Moliadze, V., Zhao, Y., Eysel, U. & Funke, K. Effect of transcranial magnetic stimulation on single-unit activity in the cat primary visual cortex. *J. Physiol.* **553**, 665–679 (2003).
48. Volianskis, A. et al. Long-term potentiation and the role of N-methyl-D-aspartate receptors. *Brain Res.* **1321**, 5–16 (2015).
49. Wang, Y. et al. Heterogeneity in the pyramidal network of the medial prefrontal cortex. *Nat. Neurosci.* **9**, 534–542 (2006).
50. Hempel, C. M., Hartman, K. H., Wang, X. J., Turrigiano, G. G. & Nelson, S. B. Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex. *J. Neurophysiol.* **83**, 3031–3041 (2000).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020

Methods

Behavioral task and recordings. *Monkey behavioral task and recordings.* Four adult (>6 years old), male rhesus monkeys (*Macaca mulatta*) were trained in an oculomotor delayed response task requiring them to fixate, view a peripheral visual stimulus on a screen at a distance of 50 cm and make a saccadic eye movement to its location after a delay period. During execution of the task, neurophysiological recordings were obtained from the dlPFC. Detailed methods of the behavioral task, training, surgeries and recordings, as well as descriptions of neuronal responses in the task, have been previously published^{6,51–54} and are only summarized briefly here. Visual stimuli were 1° squares, flashed for 500 ms at an eccentricity of either 12° or 14°, indicated as degrees of visual angle. Stimuli were randomly presented at one out of eight possible locations around the fixation point. A delay period lasting 3 s followed the presentation of the stimulus, at the end of which the fixation point turned off and a saccade terminating within 5° from the location of the remembered stimulus was reinforced with a liquid reward (5° corresponds to about 20° of arc on the circle of possible cues). Although fixation was maintained through cue and delay periods, we denote the fixation period as the interval between fixation onset and cue onset, when the only behavior expected was fixation (fixation period, Fig. 1b). A fixed ITI of 3.1 s elapsed between fixation cue extinction and the onset of the cue in the next trial (ITI, Fig. 1b). Eye position was monitored using a scleral eye coil system in two monkeys and an ISCAN camera in the other two. From two of those monkeys, we collected single-unit responses from the dlPFC using tungsten electrodes of 1–4 MΩ impedance at 1 kHz while they were performing the task⁵¹. Simultaneous recordings were obtained from arrays of 2–4 microelectrodes spaced 0.2–1 mm apart. A substantial fraction of neurons in this area showed tuned persistent delay activity during the mnemonic delay period of the task ($n = 206$ out of 822 neurons^{6,51–54}). For decoding analyses, we grouped those neurons in simultaneously recorded ensembles (total of $n = 94$ neural ensembles, 1–6 neurons per ensemble, Extended Data Fig. 1a). All experiments were conducted in accordance with the guidelines set forth by the US National Institutes of Health, as reviewed and approved by the Yale University Institutional Animal Care and Use Committee, and by the Wake Forest University Institutional Animal Care and Use Committee. Data collection and analyses were not performed blinded to the conditions of the experiments. No statistical methods were used to predetermine sample sizes, and we followed the customary practice of testing $n = 2$ monkeys for electrophysiology data and $n = 4$ monkeys for behavioral data. We note that the electrophysiology data were previously acquired and have been used in other publications^{6,51–56}.

Human participants and behavioral task. Thirty-five neurologically and psychologically healthy volunteers with normal or corrected vision (EEG experiment: $n = 15$ (4 male), 21.27 ± 4.86 years (mean \pm s.d.); two additional participants were tested, but aborted the EEG experiment with insufficient trials; TMS experiments: $n = 20$ (6 male), 29.86 years \pm 9.55 years (mean \pm s.d.); one additional participant was excluded before their MRI scan due to health concerns) from the Barcelona area provided written informed consent and were monetarily compensated for their participation, as reviewed and approved by the Research Ethics Committee of the Hospital Clínic de Barcelona. During both the EEG and TMS experiments, each participant performed two sessions lasting approximately 1.5 h. To perform behavioral and EEG analyses, we concatenated the two sessions for each participant. Stimuli were presented on a 17" HP ProBook viewed at a distance of 65 cm, and we used Psychopy (v.1.82.01) running on Python 2.7. The TMS study consisted of an initial experiment with ten participants and a preregistered replication experiment (<https://osf.io/rguzn/>) with ten more participants (Extended Data Figs. 7–9). For all three studies (one EEG and two TMS experiments), we recruited independent participant pools. For the fully randomized within-subjects design of our EEG task, condition-blind data collection and analyses were not a critical issue. In the TMS study, the experimenter could not be blinded to the location of the coil. No statistical methods were used to predetermine sample sizes, but our sample sizes were similar to those reported in relevant previous publications^{14,33,46}.

In each 1.5-h EEG session, participants completed 12 blocks of 48 trials (except for one participant, who completed 12 blocks in one session and 9 blocks in the second session). Each trial began with the presentation of a central black fixation dot (0.5 × 0.5 cm) on a gray background. After 1.1 s of fixation, a single colored circle (stimulus, diameter of 1.4 cm) appeared for 0.25 s at any of 360 circular locations at a fixed radius of 4.5 cm, randomly sampled from a uniform distribution. In 66.67% of trials (a total of 768 trials per participant), the stimulus was followed by a 1-s delay in which only the fixation dot remained visible. In the remaining trials, the delay duration was either 3 s (16.67% of trials, 192 trials per participant) or 0 s (16.67% of trials, 192 trials per participant). Trials with 0-s delay were excluded from the analyses in this study. The change in the fixation dot color (from black to the stimulus color) instructed participants to respond (response probe). Participants responded by making a mouse click at the remembered location. A transparent circle with a white border indicated the radial distance of the stimulus, so the participant was only asked to remember its angular location. After the response was given, the cursor had to be moved back to the fixation dot to self-initiate a new trial. The total length of the ITI, defined as the time between response probe and the next stimulus onset, was around 2.72 s (median,

95% confidence intervals (CIs) = [2.11 s, 4.16 s]). Participants were instructed to maintain fixation during pre-stimulus fixation, stimulus presentation and delay, and were free to move their eyes during the response and when returning the cursor to the fixation dot. Colors (one out of six colors with equal luminance) were randomly chosen with an equal probability for each trial.

Stimuli and the trial structure in the TMS task were similar to the EEG task, except for the fixation period duration (0.6 s), screen background (white), stimulus color (black) and response probe color (red). At the end of the fixation period (16.7 ms before stimulus onset), a single TMS pulse was applied in half of the vertex trials (TMS or sham trials, randomly interleaved) and in two-thirds of prefrontal trials (weak or strong TMS or sham trials, randomly interleaved). See TMS details below. Only delays of 1 s were used in this experiment. Participants completed 4 blocks of 90 (vertex) and 4 blocks of 130 (PFC) trials within each session. In the first TMS study, these eight blocks were randomly shuffled for each session. In the replication TMS study, we successively alternated vertex and PFC blocks within each session, and the two sessions of a given participant started alternately with each area in a counterbalanced design.

EEG recordings and preprocessing. We recorded EEG data from 43 electrodes attached directly to the scalp. The electrodes were located at the following modified combinatorial nomenclature sites: Fp1, Fpz, Fp2, AF7, AFz, AF8, F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, A1, T7, C5, C3, Cz, C4, C6, T8, A2, TP7, CP3, CPz, CP4, TP8, P7, P3, Pz, P4, P8, PO7, PO3, POz, PO4, PO8, O1, Oz and O2. Sites were referenced to an average of mastoids A1 and A2 and re-referenced offline to an average of all electrodes. We further recorded horizontal electrooculography data from both eyes, vertical electrooculography data from an electrode placed below the left eye and electrocardiography data to detect cardiac artifacts. We used a Brainbox EEG-1166 EEG amplifier with a 0.017–100 Hz bandpass filter and digitized the signal at 512 Hz using Deltamed Coherence software (v.5.1).

EEG data were preprocessed using Fieldtrip (v.20171231) in Matlab R2017b and R2019a. We excluded outlier trials in which variance or kurtosis across samples exceeded four standard deviations from mean variance or kurtosis over trials, respectively. To reduce artifacts in the remaining data, we ran an independent component analysis on the trial-segmented data and corrected the signal for blinks, eye movements and electrocardiogram signals, as identified by visual inspection of all components. Data were Hilbert-transformed (using the FieldTrip function `ft_freqanalysis.m`) to extract frequencies in the alpha band (8–12 Hz), and total power was calculated as the squared complex magnitude of the signal. Finally, we excluded trials in which log-normal alpha power at any electrode exceeded the time-resolved trial average of log-normal alpha power by more than four standard deviations, and trials in which the time-averaged variance across electrodes exceeded the mean variance over trials by more than four standard deviations (to increase the stability of trial-wise decoding predictions for different randomly chosen training sets). In total, we rejected an average of $3.95 \pm 1.07\%$ (mean \pm s.d.) of trials per participant. Excluding rejected trials and trials with 0-s delay, we used 914.33 ± 28.94 trials per participant. To concatenate data from the two sessions for the same participant, we normalized the alpha power of each session for each electrode separately.

TMS study. Stimulation was performed in the TMS study using a Magstim Rapid 2 machine with a 70-mm figure-of-eight coil. TMS target points were located using a BrainSight navigated brain stimulation system that allowed coordination of the coil position based on the structural MRI scan of each participant. A region of interest in the right dlPFC (MNI152 coordinates $x = 40$, $y = 34$, $z = 16$) was defined using a NeuroSynth⁵⁷ term-based meta-analysis of 53 functional MRI studies associated with the key phrase 'spatial working memory' (Supplementary Fig. 1 and Supplementary Data). This mask was transformed into the structural MRI space of each participant. Vertex target points were defined using the 10–20 measurement system. Stimulator intensity, coil position and coil orientation were held constant for each participant for the duration of each session. To mask the sound of TMS coil discharge, we had participants listen to white noise through earphones for the duration of the session. White noise volume was selected based on the threshold of the participant for detecting a TMS click using the staircase method (two up, one down). Stimulation intensity was determined by the individually defined RMT. We applied two different TMS intensities at 70% RMT (weak-TMS, 24.5–41.5% (min–max) of stimulator output) and 130% RMT (strong-TMS, 45.5–76.5% of stimulator output) depending on the trial (see main text). To reduce the number of trials per session, we applied strong-TMS at the vertex in the original study, but weak-TMS for the replication study (preregistered at <https://osf.io/rguzn/>; Extended Data Figs. 9 and 10). The stimulation parameters were in accordance with published TMS guidelines⁵⁸. In a post-experiment debriefing session, we collected information about the subjective experience of the participants. Many participants (13 out of 20) reported facial muscle twitching in the dlPFC blocks. This is an unlikely explanation for the effects observed in Fig. 6 because (1) twitching is expected to increase with TMS intensity, but we instead observed a nonlinear dependency in our effect (Fig. 6b), and (2) behavioral performance in our task as measured by the precision of the responses was not modulated by the TMS intensity in the dlPFC blocks (linear mixed model: $\theta_e^2 \sim \text{intensity} + (1|\text{subject})$, $P > 0.5$), which suggests that our

reported intensity-dependent effect (Fig. 6b) was not the result of a general behavioral impairment caused by facial twitching.

Serial bias analysis. Human study. For each trial, we measured the response error (θ_e) as the angular distance between the angle of the presented stimulus and the angle of the response. To exclude responses produced by guessing or motor imprecision, we only analyzed responses within an angular distance of 1 radian and a radial distance of 2.25 cm from the stimulus. Furthermore, we excluded trials in which the time of response initiation exceeded 3 s, and trials for which the time between the response probe of the previous trial and the stimulus presentation of the current trial exceeded 5 s. On average, $2.99 \pm 4.51\%$ (mean \pm s.d.) of trials per participant were rejected.

We measured serial biases as the average error in the current trial as a function of the circular distance between the target locations of the previous and the current trial (θ_d) in sliding windows with size $\pi/3$ and in steps of $\pi/20$ radians, and steps of $\pi/100$ radians for Fig. 2a (note that for easier interpretability, all figures depict values in angular degrees). To increase power and correct for global response biases, we calculated a ‘folded’ version of serial biases as follows²⁵. We multiplied trial-wise errors by the sign of θ_d : $\theta'_e = \theta_e \times \text{sign}(\theta_d)$, and used absolute values of θ_d . Positive mean folded errors should be interpreted as attraction toward the previous stimulus and negative mean folded errors as repulsion away from the previous location. For a scalar estimate of differences in serial bias curves (Fig. 5f), we averaged folded errors for close θ_d distances (between 0 and $\pi/2$ radians).

Monkey study. In contrast to the human study, the stimulus distribution was discrete for all the monkey experiments. On each trial, the subject was cued to one of eight possible cue locations equidistant on a circle. This restricted the minimal angular distance between cues in two consecutive trials to be $\pi/4$ radians. To obtain a finer resolution to calculate serial biases, we capitalized on the response variability on each trial: we computed θ_d as the distance between the stimulus of the current trial and the response of the previous trial (instead of the stimulus of the previous trial). Similar methods to the human study were used, except for Fig. 1a, where we used smaller sliding window sizes ($\pi/10$ in steps of $\pi/100$ radians), which was essential to capture the thinner attractive serial bias profile in monkeys (Fig. 1a). Specific differences in our monkey and human serial bias curves (Figs. 1a and 2a) may be due to the discrete stimulus distribution (eight possible locations) that we used for monkeys, in contrast to the continuous distribution used in our human experiments. Indeed, studies with larger samples and continuous stimulus distributions have reported behavioral biases in monkeys more consistent with the human literature^{20,32}. For all our serial bias curves, x axis coordinates mark the central value of the corresponding sliding window.

Statistical methods. Data were analyzed using custom scripts in Python 2.7 (monkey and TMS data) and in Python 3.7.4 (human EEG data). Details of statistical methods are tabulated in the Nature Research Reporting Summary available online. Unless stated otherwise, all hypothesis tests were two-tailed (permutation tests or bootstrap hypothesis test, $n = 10^6$) and CI are at [2.5, 97.5] percentiles of a bootstrapped distribution. Using bootstrap distributions, we avoid assuming normality for our statistical tests. One exception was the linear model used for TMS data analyses, in which normality was assumed. Supplementary Fig. 2 shows the distribution of residuals of this model and the corresponding qqplot. There was a significant deviation from normality in extreme values. This did not compromise our statistical inference because of the large sample size ($n = 18,299$ trials)⁵⁹ and because the interaction of interest was confirmed by model-free analyses (Fig. 6; Extended Data Figs. 7–9).

To test the effect of TMS on serial biases, we fit a linear mixed-effects model using the R function `lme`⁶⁰. In particular, we modeled trial-wise behavioral errors θ_e as a linear model with interaction terms for coil location (PFC versus vertex), TMS intensity (strong-TMS, sham and weak-TMS) and the sine of θ_d (prev-curr), which approximates the expected dependency of θ_e on θ_d in the presence of serial biases ($\theta_e \propto \sin(\theta_d)$). We incorporated the nonlinear dependency of serial bias on stimulation intensity that our model simulations predicted by using -1 , 0 and 1 for strong-TMS, sham and weak-TMS, respectively. In one model, we used instead the nominal percent of RMT TMS intensity used (70, 0 and 130, respectively) for comparison (Fig. 6b). We accounted for subject-by-subject variability by including random-effect intercepts and random-effect coefficients of prev-curr. The full, three-way interaction model was as follows: $\theta_e \sim \text{coil location} \times \text{intensity} \times \text{prev-curr} + (1 + \text{prev-curr} | \text{subject})$

Decoding stimulus information. Monkeys. Population decoder. For each recorded ensemble, we decoded stimulus θ_j in trial j by modeling it as a linear combination of the spike counts n_{ij} ($i = 1 \dots k$) of k simultaneously recorded neurons, computed in sliding windows of 0.5 s and steps of 0.1 s during that trial (in all decoding time courses depicted in figures (monkeys and humans), time (x axis) coordinates mark the central value of the corresponding sliding window):

$$\cos(\theta_j) \sim 1 + \sum_i^k \beta_i n_{ij} \quad \text{and} \quad \sin(\theta_j) \sim 1 + \sum_i^k \omega_i n_{ij}$$

For each set of neurons, we trained two sets of weights $\{\beta_i\}$ and $\{\omega_i\}$ on 80% of randomly selected trials and tested in the remaining trials. We applied Monte-Carlo cross-validation with 50 random splits to obtain angle estimates $\hat{\theta}_j$. We obtained a measure of error (err) by averaging across splits the mean absolute error ($|\hat{\theta}_j - \theta_j|$) in each split.

Accuracy of ensembles: distance from shuffle. To establish the significance of decoding accuracy (z), we compared the decoding error (err) for each ensemble to the distribution of decoding errors in 1,000 shuffled stimulus sequences (err_s). By shuffling the list of stimuli presented in the particular recording of each ensemble, we maintained the characteristics of the distribution (for example, unbalanced distribution of stimuli), but effectively destroyed correlations between stimuli and neural activity.

$$z = -\frac{\text{err} - \text{mean}(\text{err}_s)}{\text{s.d.}(\text{err}_s)}$$

In Fig. 1c and Extended Data Fig. 1b, we separately tested ensembles that had the strongest and weakest decoding accuracy in the delay period by obtaining z from spike counts in the delay period and classifying the ensembles based on z : ensembles within the top tertile (high-decoding delay ensembles) and those in the bottom tertile (low-decoding delay ensembles).

Accuracy of single trials: leave-one-out decoder. To measure stimulus information on a trial-by-trial basis, we used leave-one-out cross-validation (Fig. 5a–c). We regressed the β_i and ω_i weights in all trials, except the one left out for testing. For these analyses we computed spike counts in windows of 1 s in steps of 50 ms.

Humans. Linear decoder. EEG alpha power is known to decrease in occipital sites contralateral to attended locations and for locations being actively maintained in working memory^{33,61–63}. We used this feature to decode the angular position of the stimulus from the distribution of alpha power over all 43 electrodes. We trained the decoder on the stimulus label of the previous trial and decoded this information throughout the previous and current trial. Trial-wise alpha power for each electrode was modeled as a linear combination of a set of regressors representing the stimulus location in the corresponding trial, $U = WM$, where U is a $J \times K$ matrix of alpha power measured at electrode j in trial k , M is the $N \times K$ design matrix of values for regressor n in trial k , and W is the $J \times N$ weight matrix, mapping the weight for regressor n to electrode j . U and M were given by the experiment, while W was fitted using least squares.

The design matrix M is a set of eight regressors M_n representing expected ‘feature activations’⁶⁴ for feature n in trial k . The value of regressor M_n in trial k was determined as $|\sin(n\pi/8 - s_k\pi/8 + \pi/2)^2|$, where $s_k = [0 \dots 7]$ indicates which one of eight angular location bins (width $\pi/8$ radians) included the stimulus shown in trial k .

As in the monkey analyses, we measured single-trial stimulus representations using leave-one-out cross-validation, ensuring an equal number of trials from each location bin in the training set (U_i and M_i). We estimated the weight matrix \hat{W} and the design matrix \hat{M}_k for the left-out trial k , as follows:

$$\hat{W} = U_i M_i^T (M_i M_i^T)^{-1}$$

$$\hat{M}_k = (\hat{W}^T \hat{W})^{-1} \hat{W}^T U_k$$

For each trial and time point, we repeated this analysis 100 times with randomly chosen training sets (except for the temporal generalization matrix, for which ten repetitions were run, Fig. 2b), and averaged \hat{M} over all repetitions. Finally, we estimated the predicted angle $\hat{\theta}_k$ as the direction of the vector sum of feature vectors with length \hat{M}_{nk} pointing at angular location bin centers $b_n = n\pi/8$ ($n = 0 \dots 7$). Trial-wise decoding strength was then defined as $\cos(\hat{\theta}_k - \theta_k)$. To correlate the decoding strength with behavioral biases (Fig. 5d–f), we increased the stability of trial-wise measures by temporal averaging over moving 200-ms windows (x axis ticks in Fig. 5f are centered at window centers).

Cross-temporal decoding. To explore the temporal generalization of the mnemonic and the response code over time, we trained decoders in independent time windows of the previous and current trial, and tested them in all time points of consecutive trials (from 0.25 s to 1.25 s after previous stimulus onset (Fig. 2c, left), -0.25 s to 0.25 s after previous response (Fig. 2c, middle), and -1.25 s to 0.25 s after the stimulus onset of the current trial (Fig. 2c, right)). For the temporal generalization matrix (Fig. 2b), we averaged training and test data over independent windows of 50 samples (~ 97.77 ms). High-resolution time courses of mnemonic and response code (Fig. 2c) were obtained by training the decoder on averaged data from 0.5 s to 1 s after previous stimulus onset and -0.25 s to 0.25 s relative to the response time (dashed lines in Fig. 2b), respectively, and by testing on averaged data from five samples (~ 97.77 ms) through consecutive trials.

Preferred location. We computed the preferred locations of each neuron. Similar to ref. ⁶, the preferred location was determined by computing the circular mean of

the cue angles (0–315°, in steps of 45°) weighted by the mean spike count of the neuron over the delay period (3 s) following each cue presentation.

Cross-correlations. Dataset. For the estimation of functional connectivity, we estimated cross-correlations by computing the jittered cross-covariances⁶⁵ of spike counts from simultaneously recorded neuron pairs, whose preferred locations were separated by a maximum of 60° ($n = 67$). We included pairs of neurons recorded from the same electrode ($n = 21$) and pairs recorded from different electrodes ($n = 46$). For each pair, we selected those trials in which the presented cue fell within the preferred range (pref, within 40° from either preferred locations) or outside the preferred range (anti-pref, all the other trials). We discarded those trials without at least one spike for each neuron in the pair.

Jittered cross-covariance. We used the Python function `scipy.signal.correlate` to compute cross-covariances between spike trains of simultaneously recorded pairs. Spikes were counted in independent windows of 10 ms^{37,66}. For each trial, 1,000 jittered cross-covariances were computed as follows⁶⁵. We shuffled the spike counts within non-overlapping windows of 50 ms and computed cross-covariance for each of these jittered spike counts. This captured all the cross-covariance caused by slow dynamics (>50 ms) but destroyed any faster dynamics. Finally, we removed the mean of these jittered cross-covariances from the cross-covariance of each trial, ending up with correlations due to faster dynamics (≤ 50 ms). We considered the magnitude of the central peak of the cross-covariance in our analyses by averaging 3 bins (± 1 bin from the zero-lag bin). For the time-resolved cross-correlation function (Fig. 3c,d), we repeated this process for sliding windows of 1 s and steps of 50 ms, and averaged across trials and neuronal pairs.

Putative exc and inh interaction. Because changes in connectivity strength (our hypothesis for activity-silent mechanisms) affect inversely exc peaks and inh troughs of cross-correlations³⁴, we separately analyzed these two types of interactions. Similar to refs. ^{35,36}, based on the average central peak of the cross-correlation function in the entire trial [−4.5 s, 2.5 s], we classified each pair into three subgroups: (1) those with a positive peak for both pref and anti-pref trials were classified as putative exc interactions, (2) those with a negative peak for both pre and anti-pref trials were classified as putative inh interactions and (3) we discarded those with an inconsistent peak sign between pref and anti-pref trials. In total, we analyzed the cross-correlation time course of $n = 47$ pairs of neurons ($n = 27$ exc and $n = 20$ inh; from different electrodes $n = 20$ exc and $n = 13$ inh). We confirmed that our results held when analyzing only pairs from different electrodes (Fig. 3c; exc: $P = 0.01$, $n = 20$; inh: $P = 0.04$, $n = 13$, one-sided permutation test).

Delay rate versus ITI cross-correlation analyses. As shown in Fig. 3e, we sought evidence for an interplay between attractor and subthreshold network dynamics in the PFC. To this end, we computed the trial-by-trial correlation between the cross-covariance peak (see above) in the ITI—at a time point when there was no firing-rate tuning (activity-silent period, Fig. 3d)—and the mean firing rate of the two neurons at the end of the preceding delay period (last 2 s, delay-fr, Fig. 3e) for exc interaction pairs under the pref and anti-pref condition (see above). For each pair, we obtained demeaned values for each trial by subtracting the mean firing rate and the mean cross-covariance peak across all trials, respectively. This allowed us to compute the correlation based on trial-by-trial measurements of all pairs together ($n = 27$) to increase statistical power. Error bars were then computed based on a bootstrap approach on all trials for all pairs. A local activity-dependent subthreshold mechanism for ITI memory traces predicts that for pref trials, but not for anti-pref trials, firing-rate variations in the delay period determines the degree of latent variable loading (cross-covariance peak) in the ITI (Fig. 3e).

Simulating bump reactivation. We used a previously proposed computational model^{39,67,68} to study serial dependence between two consecutive trials. The model consists of a network of interconnected 2,048 excitatory and 512 inhibitory leaky integrate-and-fire neurons⁶⁹. This network was organized according to a ring structure: excitatory and inhibitory neurons were spatially distributed on a ring so that nearby neurons encoded nearby spatial locations. All connections were all-to-all and spatially tuned, so that nearby neurons with similar preferred directions had stronger than average connections, while distant neurons had weaker connections. Inhibitory-to-inhibitory connections were untuned. Network parameters were taken from ref. ⁶⁷ except for the following:

$$G_{EE, AMPA} = 0.1 \text{ nS}, G_{EI, AMPA} = 0.192 \text{ nS}$$

$$G_{EE, NMDA} = 0.42 \text{ nS}, G_{EI, NMDA} = 0.49 \text{ nS}$$

$$G_{II, GABA} = 0.7413 \text{ nS}, G_{IE, GABA} = 0.9163 \text{ nS}$$

$$g_{\text{ext}, I} = 5.8 \text{ nS}, g_{\text{ext}, E} = 5.915 \text{ nS}$$

$$J_{EE}^+ = 7.1, \sigma_{EE} = 18^\circ, J_{EI}^+ = J_{IE}^+ = 2.2, \sigma_{EI} = \sigma_{IE} = 32^\circ$$

where G values are the maximum conductances of the corresponding connections (e.g., $G_{EE, AMPA}$ is the total maximum conductance of AMPAR-mediated local excitation onto an excitatory neuron), $g_{\text{ext}, E}$ and $g_{\text{ext}, I}$ are the maximum conductance of external Poisson inputs to an excitatory or inhibitory neuron, respectively, and J^+ and σ values define the amplitude and width of corresponding connectivity footprints, respectively. See ref. ⁶⁷ for more details.

STP dynamics. Simulation of activity-silent mechanisms during the inter-trial period was done by adding two more variables x and u , as described in refs. ^{9,70}, to excitatory presynaptic neurons as follows:

$$\frac{dx}{dt} = \frac{1-x}{\tau_x} - u x \delta(t - t_{sp})$$

$$\frac{du}{dt} = \frac{U-u}{\tau_u} + U(1-u) \delta(t - t_{sp})$$

With t_{sp} marking all spike times and $\delta(t)$ being the Dirac delta function. We used the parameters $U = 0.2$, $\tau_x = 200$ ms, $\tau_u = 1,500$ ms. The effective conductance of each excitatory synapse was then $g \times u \times x$, with g being the corresponding maximum conductance parameter (see above). These STP dynamics affected only AMPA-receptor-mediated recurrent connections in the network. In a separate set of network simulations (not shown), we also included STP in inhibitory connections in the network (same parameters as indicated above) and we found that we could obtain a similar pattern of serial bias modulations as shown in Fig. 4d. This shows that our results are not specifically dependent on whether inhibitory connections present facilitation dynamics or not.

Stimulation and behavioral readout. External stimuli were fed into the circuit as weak inputs (0.25 nA) to neurons selective to the stimulus as previously described⁶⁷. Each simulation of our computational model consisted of two trials run in sequence: a first stimulus of 250 ms, a first delay period of 1,000 ms, a network resetting input (nonspecific current −0.261 nA, duration 300 ms), an ITI of 1,300 ms, a second stimulus (250 ms) and a second delay period of 1,000 ms. The first and second cue stimuli were independently drawn randomly from 360 uniformly distributed angular values, and only the network readout of the second trial was analyzed to obtain a ‘behavioral’ readout. The readout was obtained with a bump-tracking procedure: starting at cue presentation, the instantaneous network readout was derived as the angular direction of the population vector of single-neuron firing rates (computed in windows of 250 ms, sliding by 100 ms) considering the ± 100 neurons surrounding the readout estimated in the previous time step. The instantaneous readout was iteratively derived to track the center of the bump (thus ignoring possible elevated activity extending from the fixation period), and the final behavioral output was defined as the readout in the last 250 ms of the trial. Serial bias was calculated by measuring single-trial errors (behavioral readout minus target location) in relation to the angular distance θ_d between the first and second stimulus locations, as described above for experimental data.

Consecutive trials and bump reactivation. Reactivation of the previous-trial stimulus during the reactivation period (300 ms before the second stimulus onset) was accomplished by stimulating all excitatory neurons with a nonspecific external stimulus^{3,38}. This stimulus exponentially increased with a rate of $\alpha = 10 \text{ s}^{-1}$ as $\beta(1 - e^{-\alpha(t-t_0)})$, with β being the reactivation strength and t_0 the time of onset of the stimulus. The reactivation strength was weak ($\beta = 0.17$ nA) or strong ($\beta = 2.9$ nA).

Rate and synaptic tuning. For each simulation shown in Fig. 3a,b, we computed the firing rate (r) and synaptic ($s = u \times x$) tuning by computing the difference between neurons within ($\pm 50^\circ$) and outside ($180 \pm 50^\circ$) the previous bump location for both measures.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data that support the findings of this study are available at <https://github.com/comptelab/interplayPFC>.

Code availability

The custom code used in this study is publicly available at <https://github.com/comptelab/interplayPFC>.

References

- Constantinidis, C., Franowicz, M. N. & Goldman-Rakic, P. S. Coding specificity in cortical microcircuits: a multiple-electrode analysis of primate prefrontal cortex. *J. Neurosci.* **21**, 3646–3655 (2001).

52. Compte, A. et al. Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. *J. Neurophysiol.* **90**, 3441–3454 (2003).
53. Constantinidis, C., Williams, G. V. & Goldman-Rakic, P. S. A role for inhibition in shaping the temporal flow of information in prefrontal cortex. *Nat. Neurosci.* **5**, 175–180 (2002).
54. Constantinidis, C. & Goldman-Rakic, P. S. Correlated discharges among putative pyramidal neurons and interneurons in the primate prefrontal cortex. *J. Neurophysiol.* **88**, 3487–3497 (2002).
55. Murray, J. D. et al. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl Acad. Sci. USA* **114**, 394–399 (2017).
56. Wang, X. J., Tegnér, J., Constantinidis, C. & Goldman-Rakic, P. S. Division of labor among distinct subtypes of inhibitory neurons in a cortical microcircuit of working memory. *Proc. Natl Acad. Sci. USA* **101**, 1368–1373 (2004).
57. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).
58. Rossi, S., Hallett, M., Rossini, P. M., Pascual-Leone, A. & The Safety of TMS Consensus Group. Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. *Clin. Neurophysiol.* **120**, 2008–2039 (2009).
59. Lumley, T., Diehr, P., Emerson, S. & Chen, L. The importance of the normality assumption in large public health data sets. *Annu. Rev. Public Health* **23**, 151–169 (2002).
60. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-147 (2019).
61. Worden, M. S., Foxe, J. J., Wang, N. & Simpson, G. V. Anticipatory biasing of visuospatial attention indexed by retinotopically specific alpha-band electroencephalography increases over occipital cortex. *J. Neurosci.* **20**, RC63 (2000).
62. Kelly, S. P., Lalor, E. C., Reilly, R. B. & Foxe, J. J. Increases in alpha oscillatory power reflect an active retinotopic mechanism for distracter suppression during sustained visuospatial attention. *J. Neurophysiol.* **95**, 3844–3851 (2006).
63. Medendorp, W. P. et al. Oscillatory activity in human parietal and occipital cortex shows hemispheric lateralization and memory effects in a delayed double-step saccade task. *Cereb. Cortex* **17**, 2364–2374 (2007).
64. Brouwer, G. J. & Heeger, D. J. Decoding and reconstructing color from responses in human visual cortex. *J. Neurosci.* **29**, 13992–14003 (2009).
65. Amarasingham, A., Harrison, M. T., Hatsopoulos, N. G. & Geman, S. Conditional modeling and the jitter method of spike resampling. *J. Neurophysiol.* **107**, 517–531 (2012).
66. Nougaret, S. & Genovesio, A. Learning the meaning of new stimuli increases the cross-correlated activity of prefrontal neurons. *Sci. Rep.* **8**, 11680 (2018).
67. Compte, A., Brunel, N., Goldman-Rakic, P. S. & Wang, X. J. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb. Cortex* **10**, 910–923 (2000).
68. Edin, F. et al. Mechanism for top-down control of working memory capacity. *Proc. Natl Acad. Sci. USA* **106**, 6802–6807 (2009).
69. Tuckwell, H. C. *Introduction to Theoretical Neurobiology: Volume 2, Nonlinear and Stochastic Theories* (Cambridge Univ. Press, 1988).
70. Markram, H., Wang, Y. & Tsodyks, M. Differential signaling via the same axon of neocortical pyramidal neurons. *Proc. Natl Acad. Sci. USA* **95**, 5323–5328 (1998).

Acknowledgements

This work was funded by the Spanish Ministry of Science and Innovation and the European Regional Development Fund (references BFU2015-65315-R and RTI2018-094190-B-I00); by the Institute Carlos III, Spain (grant PIE 16/00014); by the Cellex Foundation; by the “La Caixa” Banking Foundation (reference LCF/PR/HR17/52150001); by the Safra Foundation; by the Generalitat de Catalunya (AGAUR 2014SGR1265 and 2017SGR01565); and by the CERCA Programme/Generalitat de Catalunya. C.C. was supported by NIH grant R01 EY017077. J.B. was supported by the Spanish Ministry of Economy and Competitiveness (FPI program, reference BES-2013-062654) and by the Bial Foundation (reference 356/18). H.S. was supported by the “La Caixa” Banking Foundation (reference LCF/BQ/IN17/11620008) and the European Union’s Horizon 2020 Marie Skłodowska-Curie grant (reference 713673). K.C.S.A. was supported by NIH grant T32-MH020002. We thank the Barcelona Supercomputing Center (BSC) for providing computing resources, and the Neurology Department of the Hospital Clinic de Barcelona for granting access to EEG, TMS and neuronavigation equipment. This work was developed at the building Centro Esther Koplowitz, Barcelona. We thank A. Morató and D. Lozano-Soldevilla for assistance with EEG analyses, L. C. García del Molino for valuable insights during the development of early versions of the model, and A. Renart and J. de la Rocha for their comments on the manuscript.

Author contributions

J.B. and A.C. performed the monkey data analyses. J.B. and A.C. developed the model. H.S. and A.C. designed the human EEG research. H.S. and A.G.-G. performed the human EEG experiments. H.S., J.B. and A.C. performed the human data analyses. A.C. and J.D. obtained the funding used for the human EEG research. K.C.S.A. performed the preliminary human EEG data analyses. J.B., R.L.M., J.V.-S. and A.C. designed the TMS experiments. R.L.M. performed the TMS experiments and performed the data analyses. S.L. performed the monkey experiments. C.C. designed the monkey research. J.B., H.S. and A.C. discussed the results and wrote the manuscript. All authors revised the manuscript and gave critical comments.

Competing interests

J.D. receives royalties from Athena Diagnostics for the use of Ma2 as an autoantibody test and from Euroimmun for the use of NMDA as an antibody test. He received a licensing fee from Euroimmun for the use of GABA_B receptor, GABA_A receptor, DPPX and IgLON5 as autoantibody tests; he has received a research grant from Sage Therapeutics.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41593-020-0644-4>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41593-020-0644-4>.

Correspondence and requests for materials should be addressed to A.C.

Peer review information *Nature Neuroscience* thanks Bradley Postle and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

3.2 Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia

In this section¹⁴, I show that serial biases in working memory are disrupted in patients with anti-NMDAR encephalitis and schizophrenia. Modeling of different NMDAR-dependent circuit alterations explains these findings most parsimoniously with a reduction in short-term synaptic potentiation¹⁵.








Supplementary material for this section is included in Appendix A2.

¹⁴ This section has been published as:

Stein, H.*, Barbosa, J.*, Rosa-Justicia, M., Prades, L., Morató, A., Galan-Gadea, A., Ariño, H. Martínez- Hernandez, E., Castro-Fornieles, J., Dalmau, J. & Compte., A. (2020). Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia. *Nat Commun*, 11, 4250. <https://doi.org/10.1038/s41467-020-18033-3>. (*equal contribution)

¹⁵ Note that in Chapters 3.2 and 3.3, I will use the abbreviation STP to refer to NMDAR-dependent short-term potentiation, in contrast with the remainder of this thesis, in which STP is used to refer to short-term plasticity.

Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia

Heike Stein ^{1,7}, Joao Barbosa ^{1,7}, Mireia Rosa-Justicia ^{1,2}, Laia Prades¹, Alba Morató¹, Adrià Galan-Gadea¹, Helena Ariño¹, Eugenia Martinez-Hernandez ^{1,3}, Josefina Castro-Fornieles ^{1,2,4}, Josep Dalmau ^{1,3,4,5,6} & Albert Compte ¹✉

A mechanistic understanding of core cognitive processes, such as working memory, is crucial to addressing psychiatric symptoms in brain disorders. We propose a combined psychophysical and biophysical account of two symptomatologically related diseases, both linked to hypofunctional NMDARs: schizophrenia and autoimmune anti-NMDAR encephalitis. We first quantified shared working memory alterations in a delayed-response task. In both patient groups, we report a markedly reduced influence of previous stimuli on working memory contents, despite preserved memory precision. We then simulated this finding with NMDAR-dependent synaptic alterations in a microcircuit model of prefrontal cortex. Changes in cortical excitation destabilized within-trial memory maintenance and could not account for disrupted serial dependence in working memory. Rather, a quantitative fit between data and simulations supports alterations of an NMDAR-dependent memory mechanism operating on longer timescales, such as short-term potentiation.

¹Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Carrer Rosselló 149, 08036 Barcelona, Spain. ²Department of Child and Adolescent Psychiatry and Psychology, 2017SGR881, CIBERSAM, Institute Clinic of Neurosciences, Hospital Clínic, Carrer Villarroel 170, 08036 Barcelona, Spain. ³Service of Neurology, Hospital Clínic, Carrer Villarroel 170, 08036 Barcelona, Spain. ⁴Department of Medicine, University of Barcelona, Carrer Casanova 143, 08036 Barcelona, Spain. ⁵Institució Catalana de Recerca i Estudis Avançats (ICREA)-IDIBAPS, Carrer Casanova 143, 08036 Barcelona, Spain. ⁶Department of Neurology, University of Pennsylvania, 3400 Spruce St, Philadelphia, PA 19104, USA. ⁷These authors contributed equally: Stein Heike, Barbosa Joao. ✉email: acompte@clinic.cat

The NMDA receptor (NMDAR) subserves memory mechanisms at several timescales, including sustained working memory delay activity^{1,2} and different temporal components of synaptic potentiation^{3–5}. In addition, hypofunction of NMDARs is linked to psychiatric disease, in particular schizophrenia⁶, and it possibly contributes to abnormal working memory function in patients with schizophrenia^{7,8}. Indeed, reduced prefrontal NMDAR density characterizes this disease⁹. Yet, the specific neural alterations by which NMDAR hypofunction could lead to memory deficits in schizophrenia are still under debate^{7,8}. Here, we studied working memory function in healthy controls, patients with schizophrenia, and patients recovering from anti-NMDAR encephalitis (see “Methods” section and Supplementary Table 1). Anti-NMDAR encephalitis is characterized by an antibody-mediated reduction of NMDARs¹⁰, accompanied by initial psychosis and long-lasting memory deficits^{11,12}. The prevalence of positive symptoms during the early stages of the disease causes frequent misdiagnosis as a schizophrenia spectrum disorder^{13,14}. Here, we tested patients that had overcome acute stages, and had progressed to a more stabilized period with some positive symptoms but dominated by negative and cognitive symptoms, comparable to those in stabilized schizophrenia patients¹⁵. Due to the parallels in neurobiology, clinical aspects, and cognition of the two diseases, we expected working memory deficits in anti-NMDAR encephalitis to qualitatively resemble those in schizophrenia. This correspondence allows linking alterations in working memory to the NMDAR in both patient groups.

We assessed memory alterations in a visuospatial delayed-response task (Fig. 1a) on two coexisting temporal scales: single-trial working memory precision as a proxy of active memory maintenance during short delays, and serial dependence of responses on previously memorized stimuli^{16,17} (serial biases, Fig. 1b) as a read-out of passive information maintenance across trials. Our results show reduced serial dependence but intact working memory precision in both patient populations. Neural correlates of this task have been identified in monkey prefrontal cortex^{18–20}, inspiring computational models that can capture key aspects of neural dynamics and behavior^{18,21,22}. The biophysical detail of these models permits to investigate how NMDAR hypofunction at different synaptic sites affects circuit dynamics and working memory. Candidate mechanisms are a disturbed balance between cortical excitation and inhibition (excitation/inhibition balance), as it is observed in schizophrenia and in studies using NMDAR antagonists (e.g., ketamine)^{26,23,24}, and alterations in NMDAR-regulated short-term synaptic potentiation^{3–5,25}. In the modeling section of this study, we systematically test the potential of these candidate mechanisms for explaining our behavioral findings. We conclude that a reduction in short-term potentiation in a network model of working memory most parsimoniously reproduces the experimentally observed memory alterations in schizophrenia and anti-NMDAR encephalitis.

Results

Unaltered working memory precision in both patients groups.

First, we sought to identify alterations in single-trial working memory precision, as an indication of a possible dysfunction of activity-based memory maintenance. Meta-analyses report mainly negative findings for delay-dependent precision impairments in schizophrenia and ketamine studies^{7,26} (but see ref. 27). We calculated the circular standard deviation of bias-corrected response errors (“Methods”) as an inverse estimate of precision for each participant and delay. Correcting for biases as a systematic source of error allowed us to estimate memory precision independently of serial biases. For all groups, precision decreased

equally with delay (Fig. 1c), indicating spared active working memory maintenance over short delays of up to 3 s in encephalitis and schizophrenia.

Patients’ memories are less biased towards previous memories.

Next, we tested whether NMDAR-related memory alterations could be observed at intermediate timescales by measuring serial dependence. Serial dependence is defined as a systematic shift of responses towards previously remembered, uncorrelated stimuli¹⁶ (Fig. 1b), revealing that traces of recently processed stimuli persist in memory circuits and are integrated with new memories. Importantly, these attractive biases emerge over the trial’s memory delay, indicating a dependence on memory processes^{28,29}. In conditions without memory requirements, only small repulsive biases are present, possibly generated during perceptual processing^{28–30}. To assess NMDAR-related differences in serial dependence, we modeled single-trial errors θ^e as a linear mixed model of delay length, group, and a non-linear basis function of the distance θ^d between consecutive stimuli^{16,29} (derivative-of-Gaussian, DoG(θ^d), “Methods”, Eq. (1); Supplementary Fig. 1), and we assessed the significance of fixed effects through ANOVA tables (“Methods”).

Serial dependence explained only a small fraction of single-trial errors in working memory (conditional $R^2 = 0.03$ for the linear model presented in Eq. (1)), reflecting its small magnitude compared to the typical extent of response inaccuracies (Fig. 1c), but it depended strongly on relevant task factors: In accordance with previous results^{28,29}, we found a dependence of attractive bias strength on memory delay (delay \times DoG(θ^d), $F(2,58) = 13.89$, $p = 1e-5$). Moreover, biases differed between groups of participants (group \times DoG(θ^d), $F(2,49) = 9.68$, $p = 0.0003$), especially when comparing groups for different delay lengths (group \times delay \times DoG(θ^d), $F(4,58) = 8.45$, $p = 2e-5$). Figure 1d–f shows linear model fits and average bias curves for 0, 1, and 3 s delays (see Supplementary Figs. 2–4 for single-subject bias curves and fits). Groupwise linear models (Eq. (2)) allowed to assess the delay dependence of biases within each population (delay \times DoG(θ^d): For healthy controls, initially repulsive biases became gradually more attractive with delay length ($F(2,17) = 26.91$, $p = 6e-6$; Supplementary Fig. 5). Encephalitis patients showed a qualitatively similar, but reduced pattern ($F(2,23) = 5.06$, $p = 0.015$). In contrast, no attractive bias emerged over delay in patients with schizophrenia ($F(2,16) = 1.31$, $p = 0.30$). Rather, a repulsive bias dominated all delay lengths in this group (DoG(θ^d), $F(1,16) = 9.07$, $p = 0.008$). Post-hoc tests and between-group comparisons are reported in Fig. 1g–i.

Serial dependence is known to fade with increasing inter-trial intervals (ITI)²⁹. We controlled for ITI length by including ITI \times DoG(θ^d) as a covariate in our linear model (“Methods”, Eq. (4); Supplementary Fig. 6): For each additional second of ITI, serial bias decreased by $0.46 \pm 0.12^\circ$ (mean \pm s.d.). However, group differences in serial dependence remained unchanged after including the covariate. The timescale of serial dependence was further defined by how many past trials influenced the current response. We observed a much weaker delay-dependent bias towards the penultimate trial, but there was no consistent evidence for group differences (Supplementary Fig. 7a–c).

Antipsychotic medication does not explain group differences.

We also controlled for potential effects of antipsychotic medication in chlorpromazine equivalents (CPZ, “Methods”) in light of significant group differences in CPZ estimates (Supplementary Table 1), and an association of CPZ with individual serial bias strength within groups (Supplementary Fig. 8). When including CPZ as a covariate (“Methods”, Eq. (5)), delay-dependent biases

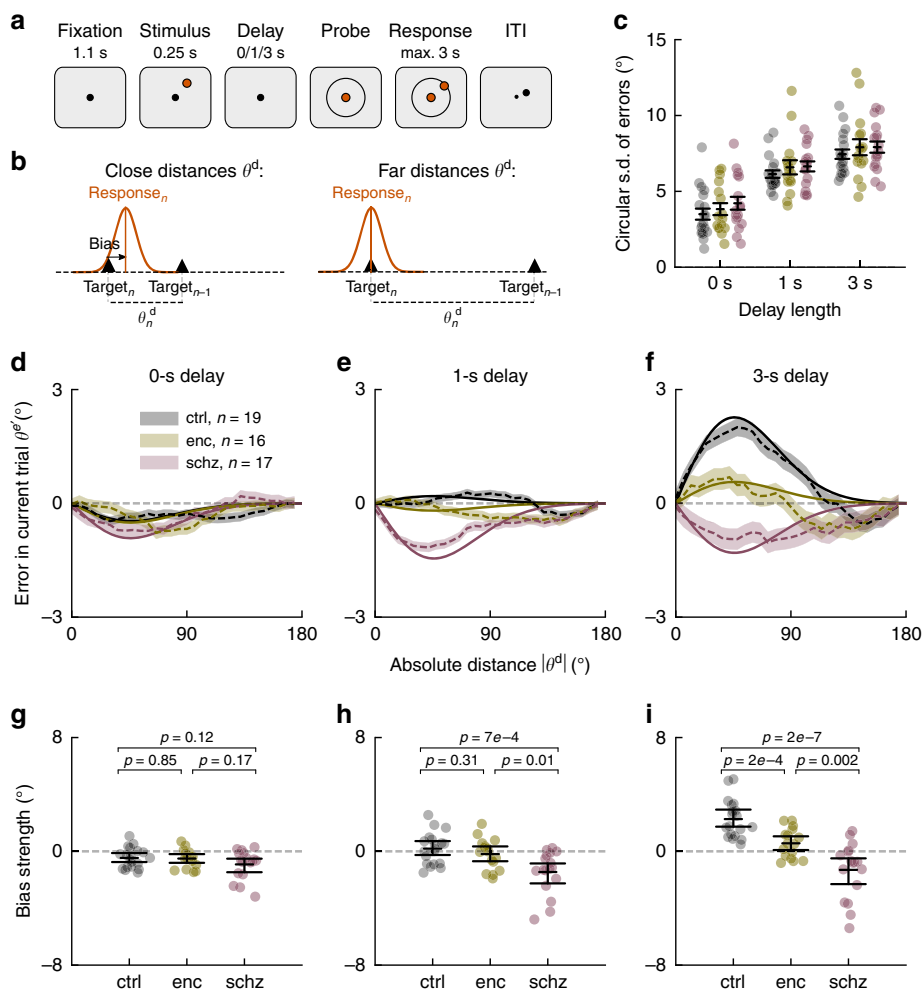


Fig. 1 **Reduced working memory-dependent serial dependence in anti-NMDAR encephalitis and schizophrenia.** **a** In each trial, subjects were to remember a stimulus that appeared for 0.25 s at a randomly chosen circular location with fixed distance from the center. Delay lengths varied randomly between trials (0, 1 or 3 s). Subjects made a mouse click to report the remembered location and started the next trial by moving the mouse back to the screen’s center during the inter-trial-interval (ITI). **b** Serial dependence is measured as a systematic shift of responses towards previous target locations. Attractive effects depend on the distance θ^d between previous and current stimulus. **c** Precision for each subject and delay was inversely estimated as the circular s.d. of bias-corrected error distributions (“Methods”). For longer delays, participants’ responses were less precise (delay, $F(2,147) = 76.87, p < 1e-16$). There were no overall or delay-dependent group differences in precision (group, $F(2,147) = 1.74, p = 0.18$; group \times delay, $F(4,147) = 0.07, p = .99$, all p -values from ANOVA). Error bars indicate \pm s.e.m. **d–f**. Serial dependence by group and delay length. Serial dependence is calculated as the ‘folded’ error θ^e for different θ^d (dashed lines; “Methods”). Solid lines show linear model fits (“Methods”), omitting intercepts and negative values of θ^d . Shading, \pm s.e.m. across pooled trials from $n = 19$ healthy controls (ctrl), $n = 17$ patients with schizophrenia (schz), and $n = 16$ patients with anti-NMDAR encephalitis (enc). **g–i** Individual (random coefficients; dots) and group estimates of serial bias strength (fixed effects; error bars indicate mean and bootstrapped 95% C.I. of the mean) by delay. **g** Serial dependence was repulsive in 0 s trials (DoG(θ^d), $F(1,52) = 12.67, p = 0.0008$), independently of group (group \times DoG(θ^d), $F(2,52) = 0.46, p = 0.63$). **h** For 1 s trials, group differences in serial dependence emerged (group \times DoG(θ^d), $F(2,48) = 6.52, p = 0.003$) between ctrl and schz ($t = 3.73, p = 7e-4$, Cohen’s $d = 1.28$) and enc and schz ($t = 2.73, p = 0.01$, Cohen’s $d = 0.98$). **i** After 3 s delay, both patient groups showed reduced biases compared to ctrl (group \times DoG(θ^d), $F(2,50) = 15.35, p = 6e-5$; ctrl vs enc, $t = 4.14, p = 2e-4$, Cohen’s $d = 1.45$; ctrl vs schz, $t = 6.44, p = 2e-7$, Cohen’s $d = 2.21$, and enc vs schz, $t = 3.40, p = 0.002$, Cohen’s $d = 1.22$). All t -tests, two-sided. In all panels, single data points show data from $n = 19$ healthy controls (ctrl), $n = 17$ patients with schizophrenia (schz), and $n = 16$ patients with anti-NMDAR encephalitis (enc).

still markedly differed between groups (Supplementary Fig. 8, caption). We designed two additional analyses to demonstrate the independence of group differences from the effect of antipsychotic medication: First, we showed that the difference in serial dependence persisted when we compared healthy controls to the unmedicated subset of encephalitis patients ($n = 12$ out of 16 encephalitis patients, Supplementary Fig. 9a–f). Second, we designed an analysis to test conservatively the group effect once we removed all the explanatory power of CPZ: We first

fitted single-trial errors θ^e as a function of CPZ and its one- and two-way interactions with delay and DoG(θ^d) in all subjects. On average, CPZ in patients with schizophrenia (370.6 ± 462.4 mg day $^{-1}$, mean \pm s.d.) explained a reduction of 1.06° in biases in the 3 s delay condition, and only a reduction of 0.08° in encephalitis patients (with CPZ equivalents of 26.6 ± 52.7 mg day $^{-1}$, mean \pm s.d.). Residuals of the linear model, now free of linear and multiplicative effects of CPZ estimates, were fitted as a function of group, delay, DoG(θ^d), and their interactions. Supplementary

Fig. 9g–l shows that group differences in memory-dependent biases remained marked (a reduction of 2.51° for schz, and 1.62° for enc in the 3 s delay condition) and highly significant even after conservatively controlling for CPZ.

Encephalitis patients' biases increase with recovery. We did not find correlations between individuals' bias estimates for 3 s delay trials and the severity of psychiatric symptoms for encephalitis or schizophrenia patients (Supplementary Fig. 8 and Supplementary Table 1). These between-subjects analyses were possibly underpowered, so we designed a within-subject longitudinal assessment for $n = 14$ encephalitis patients that returned for a follow-up session after 3–12 months (mean 8.5 months). As expected, clinical symptoms improved in these patients (Supplementary Table 2) and we found that serial dependence normalized with the patients' recovery (Eq. (8); Supplementary Fig. 10). Interestingly, for this subsample of encephalitis patients, positive and general symptoms measured in the PANSS scale correlated with serial dependence in the follow-up session (PANSS pos, $r = -0.70$, C.I. = $[-0.90, -0.26]$, $p = 0.006$; PANSS gen, $r = -0.62$, C.I. = $[-0.87, -0.13]$, $p = 0.02$), but again not significantly in the baseline session (PANSS pos, $r = -0.38$, C.I. = $[-0.76, 0.19]$, $p = 0.19$; PANSS gen, $r = -0.02$, C.I. = $[-0.54, 0.52]$, $p = 0.94$), although the direction of the effect was congruent between the two sessions. Moreover, patients with a stronger longitudinal normalization of biases improved more on the scale of positive symptoms (PANSS pos) in the follow-up session, when compared to the baseline session, $r = -0.54$, C.I. = $[-0.83, -0.02]$, $p = 0.04$ (Supplementary Fig. 10g; all correlations, Pearson's r).

Together, our experimental results show no differences in single-trial memory maintenance, but a strong reduction of delay-dependent biases in anti-NMDAR encephalitis that ameliorates with patients' recovery, and a complete absence of attractive biases in patients with schizophrenia. These findings are not explained by ITI length, general response correlations between trials (Supplementary Fig. 7d–f), response biases with respect to cardinal directions (Supplementary Fig. 11), or medication (Supplementary Fig. 8). Our conclusion is thus that alterations at the neural circuit level, related to NMDAR hypofunction, reduce serial dependence gradually, up to the point of completely disrupting attraction to previous stimuli. A prevailing idea associates NMDAR hypofunction in schizophrenia primarily to synapses onto GABAergic interneurons²³, while the role of NMDARs in working memory has been emphasized in synapses between pyramidal neurons^{1,2,21}. Alternatively, NMDARs could be involved in mechanisms directly associated with the generation of serial biases, such as short-term plasticity^{18,22,31}. To assess these mechanistic explanations comparatively, we simulated consecutive trials of a spatial working memory task in a spiking neural network model of the prefrontal cortex²¹ (Fig. 2a). Prefrontal cortex not only holds working memory contents in an activity-based code^{19,20}, but also keeps long-lasting latent (possibly synaptic) memory traces that produce serial dependence¹⁸.

NMDAR hypofunction in a prefrontal working memory circuit. We modeled a local prefrontal circuit, composed of neurons selective to the locations presented in the spatial working memory task. We used a network of excitatory and inhibitory neurons recurrently connected through AMPAR-, NMDAR- and GABA_A-mediated synaptic transmission in which persistent delay firing emerges from attractor dynamics (Fig. 2a, Supplementary Fig. 12; “Methods”). As proposed by the previous studies^{18,22,31}, we modeled serial dependence as an effect of short-

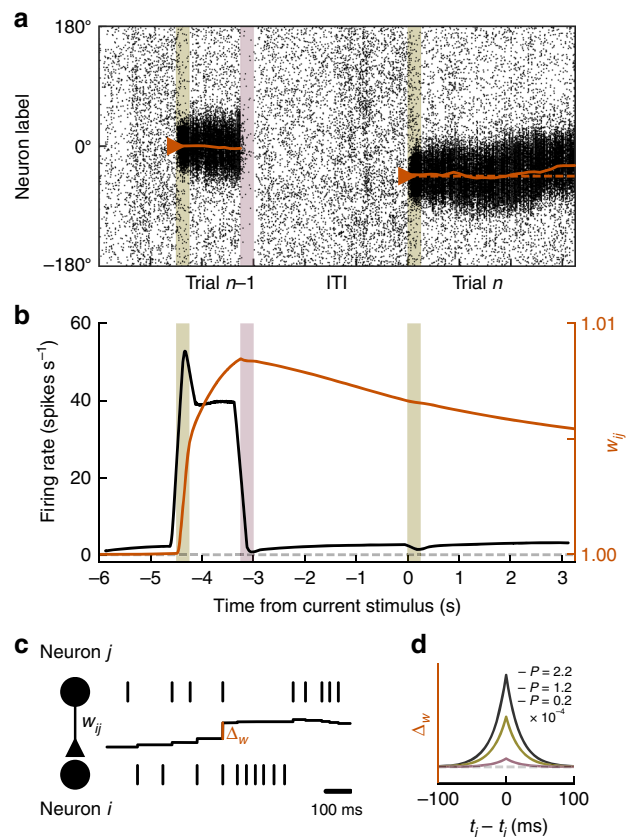


Fig. 2 Ring attractor network with synaptic STP shows serial dependence.

Simulations of two consecutive working memory trials (current trial n , previous trial $n - 1$) in a spiking neural network model with bump-attractor dynamics (“Methods”). **a** Spike times (x-axis) of excitatory neurons, ordered on y-axis by preferred angular location. Colored bars in **a**, **b** mark previous and current stimulus onset times (olive) and previous response (red). The solid orange line shows the population vector decoded from firing rates (sliding windows of 250 ms). In trial n , the active memory representation got biased towards the memory representation in trial $n - 1$. **b** Firing rate (black) and potentiated weight trace w_{ij} for neurons at 0° (orange) averaged over 1,000 trials and 20 neurons centered around 0° . Spiking activity and synaptic strength increased during trial $n - 1$ delay and decreased after the response. At current stimulus onset, information about trial $n - 1$ remained only in the potentiated weight trace. To facilitate interpretation, we excluded trials for which any neuron participated in previous and current-trial delay activity (i.e., showed firing rates >10 spikes s^{-1} after stimulus onset in trial n). **c**, **d** Associativity and decay of modeled STP. The strength of each individual synapse is determined by w_{ij} (**c**, middle black trace), which is potentiated at each spike by an amount Δ_w that depends on the relative spike times t_j and t_i of pre- and postsynaptic neurons, respectively, and on the potentiation factor P that is chosen to represent different strengths of STP (different colored lines in **d**); “Methods”, Eqs. (15) and (16)), and it is reduced by an amount relative to the synaptic strength at each presynaptic spike, resulting in activity-dependent decay (Eq. (17)).

term plasticity that builds up at delay-active recurrent excitatory synapses and maintains information during the ITI in a sub-threshold stimulus representation not reflected in firing rate selectivity (Fig. 2b, “Methods”). We implemented an associative mechanism of short-term potentiation (STP) that is NMDAR-dependent and upregulates glutamatergic efficacy, consistent with a long-lasting increase in the probability of presynaptic neurotransmitter release^{3,4}. As described in refs. 3,4, this efficacy

increase undergoes activity-dependent decay (Fig. 2c). In our simulations, stimulus-specific potentiated synaptic traces persisted through the ITI and attracted the next trial's memory representation progressively over the course of the delay^{22,31}. To mimic memory-independent repulsive biases^{29,30}, current stimulus inputs were slightly shifted away from previous stimulus values by a fixed value³¹ ("Methods"). This shift represents adaptation effects in sensory regions and is therefore not affected by local circuit alterations in prefrontal cortex.

We assessed the effects of NMDAR dysfunction on serial dependence at three potential synaptic sites: based on the reported NMDAR-dependence of STP^{3–5}, NMDAR hypofunction would reduce the strength of STP at excitatory synapses and disrupt delay-dependent biases (hypothesis I: reduced STP). Also, we tested the explanatory potential of reduced NMDAR-mediated synaptic transmission. In particular, we tested cortical disinhibition²⁷, caused by diminished NMDAR efficacy at inhibitory interneurons (hypothesis II: reduced g_{EI}), and the hypofunction of NMDARs at recurrent excitatory synapses, leading to diminished delay activity^{2,32,33} (hypothesis III: reduced g_{EE}). To assess each of these mechanisms, we independently varied STP strength, g_{EI} and g_{EE} , and we read out "behavioral responses" after 0, 1, and 3 s from population activity in our network simulations ("Methods"). Then, we fitted a linear model to measure bias strength in each condition (Eq. (18), Supplementary Fig. 13). We sought to identify which mechanisms could independently reproduce the patterns of reduced and absent biases observed in patients, and their dependence on working memory delay (Fig. 1).

Reduced STP but not E-I imbalance disrupts memory biases.

We found that both hypotheses I and III were qualitatively consistent with our experimental results: NMDAR hypofunction (whether reducing STP or g_{EE}) reduced the strength of serial dependence (Fig. 3a, c, orange). In contrast, hypothesis II was discarded by our simulations: reducing g_{EI} increased serial dependence (Fig. 3b, orange), contrary to our experimental results, and quickly led to network disinhibition, causing previous-trial delay activity to spontaneously reemerge in the ITI (Supplementary Fig. 14). Both for reduced g_{EI} and reduced g_{EE} , the percentage of outlier responses (where errors $|\theta^e| > 57.3^\circ$, i.e. 1 radian) quickly rose as the network lost the stability of one of its two states (spontaneous activity for reduced g_{EI} , and persistent delay activity for reduced g_{EE} , dashed vertical lines in Fig. 3b, c), as illustrated in Supplementary Fig. 14. Moreover, we noted that memory precision was slightly affected by all three manipulations (Fig. 3a–c), in contrast with our behavioral findings (Fig. 1b), but consistent with other studies with longer delays²⁷. Delay length and task complexity could be important factors to detect NMDAR-related differences in memory precision.

In addition, we found that hypotheses I and III could be disambiguated based on biases produced by the different linear models in 0, 1, and 3 s delays (Fig. 3d–f). Even for the lowest value of g_{EE} within the stable network regime (Fig. 3c), attractive biases increased with delay (Fig. 3f). While this manipulation can qualitatively reproduce decreased delay-dependent biases in the encephalitis group, it is incompatible with our results for patients with schizophrenia (Fig. 1), who do not develop attractive biases in memory trials. In contrast, reduced STP at recurrent excitatory synapses captured a pattern of equally strong repulsive biases for all delay lengths (Fig. 3d). Note that these findings also hold for a network with STP (and NMDAR-dependent reductions in STP) in inhibitory interneurons³⁴ (Supplementary Fig. 15). Based on our simulations, we conclude that the disruption of STP, a mechanism operating on a longer timescale than activity-based

memory maintenance, provides a plausible explanation for altered serial dependence as observed in schizophrenia and anti-NMDAR encephalitis.

Discussion

In this study, we assessed working memory alterations in two patient groups linked to NMDAR hypofunction, and hypothesized that their shared clinical and neurobiological features should be reflected in qualitatively similar behavioral patterns. In accordance with this reasoning, we found a drastic reduction of working memory serial dependence both in patients with anti-NMDAR encephalitis and schizophrenia, as compared to healthy controls. In contrast, we did not find memory maintenance deficits on timescales of a few seconds, suggesting that cognitive deficits in these patients^{8,12} might be partly explained by the disruption of long-lasting, inactive memory traces, and a lacking integration of past and current memories. Our modeling results show that simple alterations in cortical excitation (hypotheses II and III), as proposed by current theories of NMDAR hypofunction in schizophrenia^{6,24,27}, cannot fully explain these behavioral findings. Instead, altered serial dependence is mechanistically accounted for by a disruption in slower dynamics, here specified as NMDAR-dependent associative STP (hypothesis I) that is triggered by sustained delay activity and influences memory representations in upcoming trials. Our results suggest that clinical reports of short-term memory alterations in schizophrenia and anti-NMDAR encephalitis could be understood in the light of reduced synaptic potentiation²⁵. This is consistent with *in vitro* studies, which have demonstrated the dependence of STP on specific subunit components of the NMDAR^{3,4}, and reduced STP in genetic mouse models of schizophrenia³⁵. Importantly, our modeling is not incompatible with altered cortical excitatory or inhibitory tone as a result of hypofunctional NMDARs. Rather, it states the necessity of assuming alterations in a mechanism operating on longer timescales, such as STP. For instance, diminished STP alongside symmetric effects on both E-E and E-I synapses could maintain the excitation/inhibition balance and thus stable delay activity, while interrupting passive between-trial information maintenance.

Future studies should address the effects of pharmacological NMDAR blockade on serial dependence. These studies could unequivocally confirm the role of the NMDAR for trial-history effects in working memory, and at the same time allow to ask more specific questions: On the one hand, serial dependence effects under different NMDAR antagonists should vary according to how blocking specific NMDAR subunits modulates synaptic potentiation at different timescales³. Our results cannot address subunit specificity because anti-NMDAR encephalitis (and possibly schizophrenia⁹) is associated with hypofunction of the GluN1 subunit, which is contained in all NMDARs^{36,37}. On the other hand, pharmacological studies in combination with neural recordings could reveal how trial-history representations are affected by the blockade of NMDARs^{18,38}. In rodents, long-term pharmacological experiments during behavior could be complemented with *in vitro* studies to assess STP directly. Finally, pharmacological studies would clarify if the alterations in serial dependence occur as a result of acute NMDAR hypofunction or whether they depend on compensatory changes in STP that arise after early, acute phases of cortical excitation/inhibition imbalance in these diseases (e.g., as a long-term adjustment of the probability of presynaptic neurotransmitter release).

We showed how working memory in the two investigated diseases is altered in a parallel way, and how these alterations are parsimoniously explained by manipulating a single, NMDAR-dependent synaptic variable in our model. However, substantial

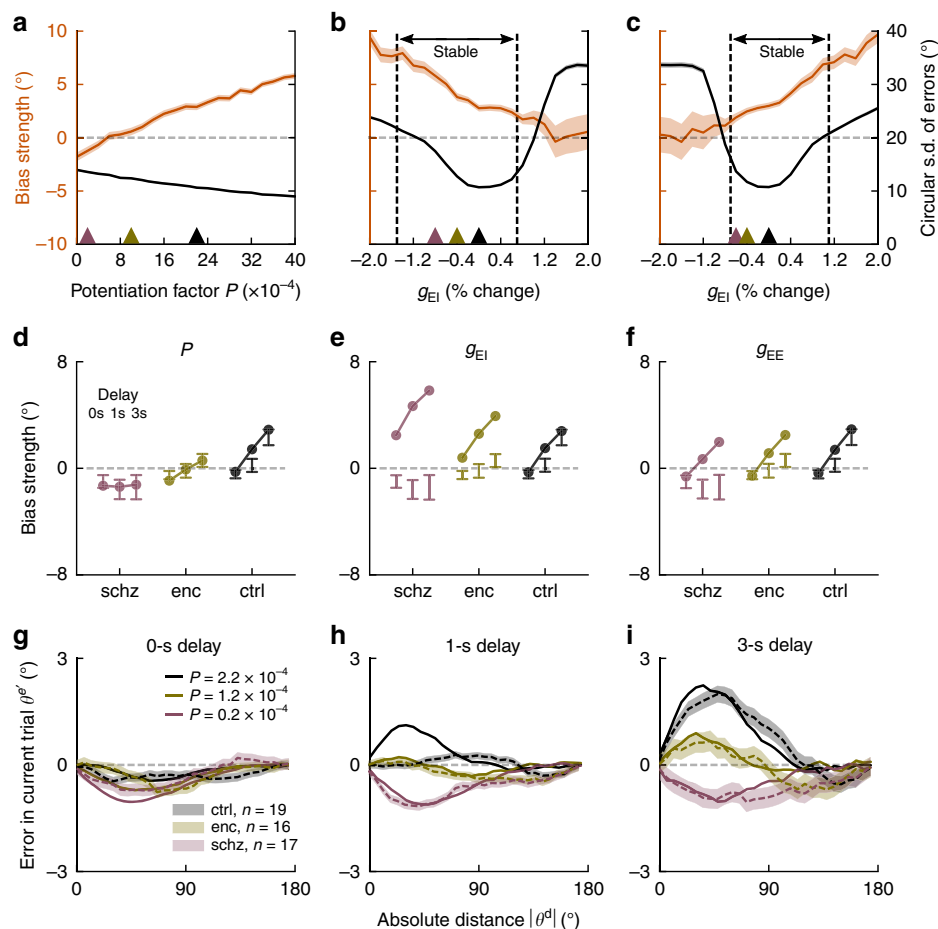


Fig. 3 Altered STP simulates reduced serial dependence in spiking neural networks. **a–c** Serial dependence (orange, bias coefficients from linear model, “Methods”) and precision (black, circular s.d. of errors) as a function of model parameters in 3 s delay trials (20,000 trials per parameter value). Vertical dashed lines indicate transition to unstable network regimes for which more than 10% of trials were outliers ($|\theta^d| > 57.3^\circ$, i.e., 1 radian). Shading, 95% C.I. for regression estimates of bias coefficients in simulated responses. **a** Serial dependence decreased gradually when decreasing STP (potentiation factor P), while the network remained stable for all simulated values of P . Precision changed slightly as a function of STP. **b** Cortical disinhibition via decreased g_{EI} augmented serial bias while strongly affecting precision and stability, either due to instability of persistent activity (right, Supplementary Fig. 14b), or due to instability of spontaneous activity (left, Supplementary Fig. 14a). **c** Lowering recurrent cortical excitation (g_{EE}) led to the opposite pattern, decreasing biases. **d–f** Delay dependence of biases for each group, as defined by parameter values in (a–c), (respectively colored triangles). Points depict mean bias strength (over 20,000 trials) for each parameter value. For comparison, error bars indicate 95% CI for bias strength obtained from $n = 19$ healthy controls (ctrl), $n = 17$ patients with schizophrenia (schz), and $n = 16$ patients with anti-NMDAR encephalitis (enc) (reordered from Fig. 1g–i). **d** Lowering STP strength reproduced the experimental data. In **e**, **f** reduction of NMDAR conductances (g_{EI} or g_{EE}) did not reproduce group and delay dependencies of experimental biases. **g–i** Solid lines, simulated serial dependence by delay length for different values of P , indicated by colored triangles in (a) (20,000 trials per potentiation level P). Dashed lines with error bars, serial dependence in encephalitis, schizophrenia, and healthy controls. Bias calculated as averaged ‘folded’ error θ^d for binned absolute previous-current distances θ^d . Shading, \pm s.e.m. Compare to Supplementary Fig. 15 for a network with STP (and STP disruptions in patients) in both E–E and E–I connections.

neurobiological heterogeneity must underlie the differences in epidemiology and longitudinal development of schizophrenia and autoimmune anti-NMDAR encephalitis³⁹. Under this reasoning, we cannot exclude that distinct biological mechanisms in our two patient groups might lead to convergent patterns of working memory processing. For instance, our modeling shows that encephalitis patients’ biases could also be explained qualitatively by a reduced excitation-to-inhibition ratio in the memory circuit (Fig. 3f), consistent with task-related fMRI BOLD activity in ketamine³³, and the effect of NMDAR antagonists on single-cell firing rates in monkey PFC². In contrast, we could not confirm the findings of previous modeling work of schizophrenia, postulating that deficits in working memory precision and higher susceptibility to distractors^{40,41} or alterations in probabilistic reasoning⁴² could be explained by an increased excitation-to-

inhibition ratio, leading to cortical disinhibition. This mechanistic alteration cannot replicate serial dependence deficits in schizophrenia in our model (Fig. 3b, e). Reduced short-term plasticity, in contrast, would predict reduced working memory precision after long memory delays (Fig. 3a, see also ref. 43), and higher susceptibility to distractors⁴⁴ in line with reported behavior in schizophrenia⁴¹, which was previously proposed to reflect an excessive excitation-to-inhibition ratio. In addition, some incongruities with previous findings might be explained by the acuteness of the patients’ condition, with more acute or psychotic stages being connected with patterns of disinhibition, and less acute stages with residual alterations in synaptic plasticity, but not cortical excitation. Alternatively, mechanisms not considered in our model could be at play. For instance, NMDAR dysfunction could negatively affect long-range connectivity^{45–47} between trial

history-tracking areas³⁸ and areas that hold current working memory contents (like prefrontal cortex), and in this way impede the integration of previous with current memories. Note, however, that recent combined experimental and theoretical work in primate and human prefrontal cortex shows how both past and current memories are jointly represented in prefrontal cortex, and how their interaction subserves serial dependence¹⁸.

Our findings advance the conceptual understanding of working memory alterations in schizophrenia and anti-NMDAR encephalitis, as they demonstrate a selective disruption of information carryover between trials, reflected by a reduction of serial dependence that is robustly found in neurotypical subjects¹⁷. We found several indicators of clinical relevance for our finding. First, as anti-NMDAR encephalitis patients recovered, their biases normalized in the direction of healthy controls (Supplementary Fig. 10a–c). Second, the amount of this normalization correlated across patients with their improvement on a scale that measures positive symptoms (Supplementary Fig. 10g), indicating a potential relation between psychotic symptoms and reductions in serial dependence. Third, both the alterations in serial dependence and the strength of positive symptoms were higher for patients with schizophrenia than for the anti-NMDAR encephalitis group. Still, studies with larger sample sizes are needed to confirm the relation of psychotic symptoms and reduced serial biases at the subject-level, which in our study did not reach significance for two out of three analyses in patients with schizophrenia and anti-NMDAR encephalitis (Supplementary Fig. 8 and “Results”).

Serial dependence could also reflect a clinically relevant dimension which is not or only mildly related to the assessed psychiatric scales. In this sense, it has been argued that serial dependence could facilitate information processing in temporally coherent real-world situations¹⁷. Alternatively, serial biases could be the mere by-product of long-lasting cellular or synaptic mechanisms that support memory stabilization during working memory delays⁴⁸. Our study is in line with previous findings of reduced susceptibility to proactive interference in schizophrenia^{49,50}. However, while proactive interference is mainly discussed in the context of cognitive control, the limited complexity of our task restricts possible interpretations of reduced between-trial interference and supports the role of reduced residual memory traces. Moreover, thanks to our task’s well-studied single-neuron correlates^{18–20} and biophysical models^{18,19,21} and the comparison with anti-NMDAR encephalitis patients, we provide a specific mechanistic model of synaptic deficits leading to reduced previous-trial interference in schizophrenia.

Interestingly, a reduction in serial dependence has recently been reported for patients with autism⁵¹, a disease also associated with NMDAR hypofunction⁵² and alterations in synaptic potentiation²⁵. Further, as for autism, our findings of reduced serial dependence are compatible with normative accounts of information processing in schizophrenia. Classic theories and recent studies have reported an underweighting of past context, or in Bayesian terms, learned priors, and an overweighting of incoming perceptual information in patients with schizophrenia^{42,53,54} and NMDAR hypofunction⁵⁵. Long-lived traces of past stimuli could serve as Bayesian priors to perception and memory, and a disruption of STP might be regarded as a biological implementation of a reduced usage of priors in schizophrenia and anti-NMDAR encephalitis.

Methods

Experimental sample. We included $n = 16$ patients with anti-NMDAR encephalitis (enc), $n = 17$ patients with schizophrenia or schizoaffective disorder ($n = 12$ and $n = 5$, respectively; schz), and $n = 19$ neurologically and psychiatrically

healthy control participants (ctrl), all with normal or corrected vision. Behavioral data from $n = 14$ healthy controls has been included in a previous study¹⁸. Psychiatric diagnoses (or the absence thereof for controls) were confirmed using the Structured Clinical Interview for DSM IV (SCID-I)⁵⁶. Patients diagnosed with anti-NMDAR encephalitis were recruited from different centers ($n = 14$ in Spain, $n = 1$ in Germany and $n = 1$ in the United Kingdom) at the moment of hospital discharge and completed the experiment around 5.5 months after disease onset (median, interquartile range i.q.r. = 3.7–7.2 months). All patients fulfilled clinical diagnostic criteria of anti-NMDAR encephalitis with confirmation of CSF IgG antibodies against the GluN1 subunit of the NMDAR⁵⁷. All subjects were tested in our laboratory for antibodies against NMDAR in serum³⁶ and all healthy controls and patients with schizophrenia were seronegative. Anti-NMDAR encephalitis is known to have a prolonged process of recovery after the acute stage of the disease⁵⁸, and patients in the prolonged recovery phase still suffer from cognitive deficits as has been previously described in cohorts with long follow-up¹². All patients were sufficiently recovered to participate in the testing procedure. Controls and patients with schizophrenia were recruited from the Barcelona area and from Hospital Clínic (Barcelona, Spain), respectively. Patients with schizophrenia were tested 35.0 months after diagnosis (median, i.q.r. = 16.0–69.5 months) and were clinically stable at the time of testing. All participants (and, in the case of minors of age, their legal guardians) provided written informed consent and were monetarily compensated for their time and travel expenses, as reviewed and approved by the Research Ethics Committee of Hospital Clínic. All subjects were assessed for psychiatric symptoms and functionality through a battery of standard tests including the Spanish versions of the Positive and Negative Syndrome Scale (PANSS)⁵⁹, the Young Mania Rating Scale (YMRS)⁶⁰, the Hamilton Depression Rating Scale (HAM-D)⁶¹ and the Global Assessment of Functioning Scale (GAF)⁶². Finally, the dose of antipsychotic medication at the moment of testing was estimated as chlorpromazine equivalent (CPZ, mg day⁻¹)⁶³. For a demographic and clinical overview of the populations, please refer to Supplementary Table 1.

Experimental task protocol and behavioral testing. Participants completed two 1.5 h sessions performing a visuospatial working memory task described in Fig. 1a. In each session, participants were asked to complete 12 blocks of 48 trials. However, some participants did not complete all blocks (on average, participants completed 1114.1 ± 134.4 trials (mean \pm s.d., ctrl), 1086.0 ± 189.9 trials (enc), and 1030.6 ± 192.8 trials (schz)).

For stimulus presentation, we used Psychopy v3.1.5 on Python 2.7, running on a 17" HP ProBook laptop. Each trial began with the presentation of a central black fixation square on a gray background (0.5×0.5 cm) for 1.1 s. A single colored circle (stimulus, diameter 1.4 cm, 1 out of 6 randomly chosen colors with equal luminance) was then presented during 0.25 s at one of 360 randomly chosen angular locations at a fixed radius of 4.5 cm from the center. The stimulus was followed by a randomly chosen delay of 0 (16.67% of trials), 1 (66.67% of trials), or 3 s (16.67% of trials) in which only the fixation dot remained visible (except for 0 s trials, where the stimulus remained visible until the participant started to move the cursor). When the fixation dot changed to the stimulus’ color (probe), participants were asked to respond by making a mouse click at the remembered location (response). A white circle indicated the stimulus’ radial distance, so participants only had to remember the angular position. After the response, the cursor had to be moved back to the fixation dot to start a new trial (ITI). Participants were instructed to maintain fixation during the fixation period, stimulus presentation, and memory delay and were free to move their eyes during response and when returning the cursor to the fixation dot.

Error and serial dependence analysis. Response errors θ_n^e in trial n were measured as the angular distance between response and target. To exclude errors due to guessing or motor imprecision, we only analyzed responses within an angular distance of 1 radian and a radial distance of 2.25 cm from the stimulus. Further, we excluded trials in which the time of response initiation exceeded 3 s, and trials for which the time between the previous trial’s response probe and the current trial’s stimulus presentation exceeded 5 s. In total, $2.6 \pm 4.2\%$ (mean \pm s.d., ctrl), $4.8 \pm 6.9\%$ (enc) and $7.5 \pm 9.6\%$ (schz) of trials per participant were rejected (but only $0.1 \pm 0.2\%$ (ctrl), $0.4 \pm 0.5\%$ (enc) and $0.6 \pm 0.7\%$ (schz) of trials were excluded due to angular response errors).

We then measured serial dependence as the error in the current trial as a function of the circular distance between the previous and the current trial’s target location. Figure 1c–e depict ‘folded’ serial dependence: We multiplied trial-wise errors θ_n^e by the sign of the previous-current distance, $\theta_n^d: \theta_n^e = \theta_n^e * \text{sign}(\theta_n^d)$, and then binned data based on absolute values $|\theta_n^d|$. Errors θ_n^e were then averaged for each $|\theta_n^d|$ in sliding windows with size $\pi/3$ in steps of $\pi/30$. Positive mean folded errors should be interpreted as attraction towards the previous stimulus and negative mean folded errors as repulsion away from the previous location. In all figures including bias curves, s.e.m. are calculated across pooled trials from all subjects for each group and delay. For visualization, all values were transformed from radians to angular degrees.

Linear (mixed) models. We modeled signed errors θ_{nm}^e in trial n and subject m using linear mixed models that included the dummy-coded variables group (ctrl,

enc or schz) and delay (0, 1, or 3 s), and a nonlinear function of previous-current stimulus distance θ_{nm}^d , DoG(θ_{nm}^d), which has been used for modeling serial dependence^{16,29}. DoG(θ_{nm}^d) is the normalized first derivative of a Gaussian with fixed location hyperparameter $\mu = 0$. Its scale parameter σ was determined using cross-validation as explained below (see also Supplementary Fig. 1). Our main linear model is:

$$\begin{aligned} \theta_{nm}^e = & \beta_0 + \beta_{1,g} \text{group}_{nm} + \beta_{2,d} \text{delay}_{nm} - \beta_3 \text{DoG}(\theta_{nm}^d) \\ & + \beta_{4,g,d} \text{group}_{nm} \text{delay}_{nm} - \beta_{5,g} \text{group}_{nm} \text{DoG}(\theta_{nm}^d) \\ & - \beta_{6,d} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) - \beta_{7,g,d} \text{group}_{nm} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) \\ & + \gamma_{0,m} - \gamma_{1,m} \text{DoG}(\theta_{nm}^d) - \gamma_{2,m,d} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) + \epsilon_{nm} \end{aligned} \quad (1)$$

β coefficients estimate fixed, and γ coefficients random effects. Coefficient subscripts g and d denote that a separate coefficient was estimated for different values of dummy-coded variables group or delay, respectively, resulting in a total of 18 β coefficients for Eq. (1). Coefficient subscript m denotes that a separate coefficient was estimated for each subject. Bias strength for a certain condition can then be read out as the sum of coefficients of all terms containing DoG(θ_{nm}^d) and the dependence of bias strength on other variables is assessed by evaluating the significance of interaction terms containing DoG(θ_{nm}^d) and the relevant variable. To measure response precision, bias-corrected response errors were defined as linear model residuals ϵ_{nm} from Eq. (1). For each subject and delay, inverse response precision was then measured as the circular s.d. of ϵ_{nm} .

Group- (Eq. (2), Supplementary Fig. 5) and delay-wise (Eq. (3), Fig. 1g–i) linear models were defined as:

$$\begin{aligned} \theta_{nm}^e = & \beta_0 + \beta_{1,d} \text{delay}_{nm} - \beta_2 \text{DoG}(\theta_{nm}^d) - \beta_{3,d} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) \\ & + \gamma_{0,m} - \gamma_{1,m} \text{DoG}(\theta_{nm}^d) - \gamma_{2,m,d} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) + \epsilon_{nm} \end{aligned} \quad (2)$$

$$\begin{aligned} \theta_{nm}^e = & \beta_0 + \beta_{1,g} \text{group}_{nm} - \beta_2 \text{DoG}(\theta_{nm}^d) - \beta_{3,g} \text{group}_{nm} \text{DoG}(\theta_{nm}^d) \\ & + \gamma_{0,m} - \gamma_{1,m} \text{DoG}(\theta_{nm}^d) + \epsilon_{nm} \end{aligned} \quad (3)$$

The effect of covariates ITI length (Eq. (4)) and CPZ equivalent (Eq. (5)) were assessed as:

$$\begin{aligned} \theta_{nm}^e = & \beta_0 + \beta_{1,g} \text{group}_{nm} + \beta_{2,d} \text{delay}_{nm} - \beta_3 \text{DoG}(\theta_{nm}^d) \\ & + \beta_{4,g,d} \text{group}_{nm} \text{delay}_{nm} - \beta_{5,g} \text{group}_{nm} \text{DoG}(\theta_{nm}^d) \\ & - \beta_{6,d} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) - \beta_{7,g,d} \text{group}_{nm} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) \\ & - \beta_8 \text{ITI}_{nm} \text{DoG}(\theta_{nm}^d) + \gamma_{0,m} - \gamma_{1,m} \text{DoG}(\theta_{nm}^d) \\ & - \gamma_{2,m,d} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) + \epsilon_{nm} \end{aligned} \quad (4)$$

$$\begin{aligned} \theta_{nm}^e = & \beta_0 + \beta_{1,g} \text{group}_{nm} + \beta_{2,d} \text{delay}_{nm} - \beta_3 \text{DoG}(\theta_{nm}^d) \\ & + \beta_{4,g,d} \text{group}_{nm} \text{delay}_{nm} - \beta_{5,g} \text{group}_{nm} \text{DoG}(\theta_{nm}^d) \\ & - \beta_{6,d} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) - \beta_{7,g,d} \text{group}_{nm} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) \\ & - \beta_{8,d} \text{CPZ}_{nm} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) + \gamma_{0,m} - \gamma_{1,m} \text{DoG}(\theta_{nm}^d) \\ & - \gamma_{2,m,d} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) + \epsilon_{nm} \end{aligned} \quad (5)$$

Further, a conservative estimate of group effects when controlling for CPZ equivalents was obtained by first regressing trialwise errors as CPZ-dependent effects excluding random effects to not absorb variance related to the experimental group that subjects belonged to (notice dropped m subscripts):

$$\begin{aligned} \theta_n^e = & \beta_0 + \beta_1 \text{CPZ}_n + \beta_{2,d} \text{CPZ}_n \text{delay}_n - \beta_3 \text{CPZ}_n \text{DoG}(\theta_n^d) \\ & - \beta_{4,d} \text{CPZ}_n \text{delay}_n \text{DoG}(\theta_n^d) + \epsilon_n \end{aligned} \quad (6)$$

and subsequently modeling residuals ϵ_n as main and interaction effects of group, delay, and DoG(θ_{nm}^d) as described in Eq. (1) (Supplementary Fig. 9g–l).

Biases towards stimuli in trial $n - 2$ were measured by including distances to the penultimate stimulus, $\theta_{nm}^{d'}$,

$$\begin{aligned} \theta_{nm}^e = & \beta_0 + \beta_{1,g} \text{group}_{nm} + \beta_{2,d} \text{delay}_{nm} - \beta_3 \text{DoG}(\theta_{nm}^d) \\ & + \beta_{4,g,d} \text{group}_{nm} \text{delay}_{nm} - \beta_{5,g} \text{group}_{nm} \text{DoG}(\theta_{nm}^d) \\ & - \beta_{6,d} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) - \beta_{7,g,d} \text{group}_{nm} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) \\ & - \beta_8 \text{DoG}(\theta_{nm}^{d'}) - \beta_{9,g} \text{group}_{nm} \text{DoG}(\theta_{nm}^{d'}) - \beta_{10,d} \text{delay}_{nm} \text{DoG}(\theta_{nm}^{d'}) \\ & - \beta_{11,g,d} \text{group}_{nm} \text{delay}_{nm} \text{DoG}(\theta_{nm}^{d'}) + \gamma_{0,m} - \gamma_{1,m} \text{DoG}(\theta_{nm}^d) \\ & - \gamma_{2,m,d} \text{delay}_{nm} \text{DoG}(\theta_{nm}^d) + \epsilon_{nm} \end{aligned} \quad (7)$$

Baseline and follow-up sessions in encephalitis patients and controls were compared by:

$$\begin{aligned} \theta_n^e = & \beta_0 + \beta_{1,\text{session}_n} + \beta_{2,g} \text{group}_n + \beta_{3,d} \text{delay}_n \\ & - \beta_4 \text{DoG}(\theta_n^d) + \beta_{5,g} \text{session}_n \text{group}_n + \beta_{6,d} \text{session}_n \text{delay}_n \\ & + \beta_{7,g,d} \text{group}_n \text{delay}_n - \beta_8 \text{session}_n \text{DoG}(\theta_n^d) \\ & - \beta_{9,g} \text{group}_n \text{DoG}(\theta_n^d) - \beta_{10,d} \text{delay}_n \text{DoG}(\theta_n^d) \\ & - \beta_{11,g} \text{session}_n \text{group}_n \text{DoG}(\theta_n^d) - \beta_{12,d} \text{session}_n \text{delay}_n \text{DoG}(\theta_n^d) \\ & - \beta_{13,g,d} \text{group}_n \text{delay}_n \text{DoG}(\theta_n^d) - \beta_{14,g,d} \text{session}_n \text{group}_n \text{delay}_n \text{DoG}(\theta_n^d) + \epsilon_n \end{aligned} \quad (8)$$

where session_n takes values 0 or 1 (baseline vs. follow-up). In this model, we did not include random effects due to increased model complexity and resulting difficulties in model convergence. For extended linear models in Eqs. (4), (5), (7), and (8), we compared nested models via Wald Tests to determine the optimal model complexity. Data was analyzed in Python 3.7. We used different packages from R statistics (version 3.6.3) through the ‘rpy2’ interface⁶⁴. All linear mixed models were fitted, compared and statistically tested with packages ‘lme4’⁶⁵ and ‘lmerTest’⁶⁶, which calculates ANOVA tables for the fixed effects of the linear mixed model by estimating degrees of freedom and F values using Satterthwaite’s method. For optimization, we used the ‘optimx’ package⁶⁷ ‘nlmb’ algorithm with a convergence tolerance of 0.003 and checked the consistency of parameter estimates with other optimization algorithms (‘L-BFGS-B’, ‘bobyqa’). Note that the normality assumption of residuals was not met (normality test, $s^2 + k^2 = 4248.72$, $p < 1e-16$), but with only slightly diverting kurtosis (Fisher) = 3.37 and skewness = 0.12 parameters. Due to the large number of trials ($n = 52,394$), this should not compromise statistical inference⁶⁸. Moreover, all effects of relevant task variables are visualized both in a model-based and model-free way to confirm their congruence.

Basis function selection and hyperparameter cross-validation. To determine the hyperparameter σ used in Eqs. (1)–(8), we fitted errors θ_n^e in trial n as a linear model including factors group, delay, and DoG(θ_n^d) as described in Eq. (1), but excluding random effects:

$$\begin{aligned} \theta_n^e = & \beta_0 + \beta_{1,g} \text{group}_n + \beta_{2,d} \text{delay}_n - \beta_3 \text{DoG}(\theta_n^d) \\ & + \beta_{4,g,d} \text{group}_n \text{delay}_n - \beta_{5,g} \text{group}_n \text{DoG}(\theta_n^d) \\ & - \beta_{6,d} \text{delay}_n \text{DoG}(\theta_n^d) - \beta_{7,g,d} \text{group}_n \text{delay}_n \text{DoG}(\theta_n^d) + \epsilon_n \end{aligned} \quad (9)$$

while setting Gaussian hyperparameters $\mu = 0$ and $\sigma \in [0.2, 1.8]$ (in radians). For each value of the scale parameter σ , we used a stratified cross-validation procedure, fitting the linear model to 67% of the trials from each subject and testing the prediction in the left-out 33% of trials. Performance for each σ was evaluated using the mean squared error (MSE) of predictions from 1000 cross-validation repetitions. σ was chosen so as to minimize the MSE obtained by the linear model, yielding $\sigma = 0.8$ (Supplementary Fig. 1).

To test whether a linear model with repulsive biases at high distances $|\theta_n^d|$ fitted our data more parsimoniously, we compared cross-validation MSE for linear models with first- and third-derivative-of-Gaussian basis functions (Supplementary Fig. 1). We repeated the hyperparameter fitting procedure described above for the third-derivative-of-Gaussian model using hyperparameters $\mu = 0$ and $\sigma \in [0.6, 2.0]$ rad. As the first-derivative-of-Gaussian model produced smaller MSE in the cross-validation procedure, we discarded the third-derivative-of-Gaussian model. Thus, all linear model results reported in this manuscript correspond to the first-derivative-of-Gaussian model.

Confidence intervals and effect sizes. We compared single-subject bias estimates between groups using post hoc t -tests. Effect sizes for these comparisons were estimated as Cohen’s d , defined as $d = \frac{\mu_1 - \mu_2}{s}$ for independent samples, where s is

the pooled standard deviation: $s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$, and as $d = \frac{t}{\sqrt{n}}$ for related samples. For correlations of individual subjects’ biases with symptoms, we used Pearson correlation and calculated parametric 95% confidence intervals (‘CI’ function from the ‘psychometric’⁶⁹ package). In the face of small, potentially non-normal samples, we confirmed significant results with bootstrap confidence intervals and p -values, leading to consistent results in all but one correlation (Supplementary Fig. 10g): Here, we obtained C.I. = $[-0.83, -0.02]$ and $p = 0.04$ with parametric methods, but C.I. = $[-0.85, 0.09]$ and $p = 0.09$ with non-parametric methods (all two-sided; note however that our directed hypothesis of an expected negative correlation supports a one-sided test with $p = 0.04$). Confidence intervals of the mean (Figs. 1 and 3, and Supplementary Figs. 5, 6 and 9) were calculated as 95% bootstrap confidence intervals.

Neural network architecture and dynamics. We simulated consecutive pairs of trials in a spiking neural network model of prefrontal cortex implemented in Brian2⁷⁰. $N_E = 1024$ excitatory and $N_I = 256$ inhibitory leaky integrate-and-fire

neurons were connected all-to-all via synapses governed by NMDAR-, AMPAR-, and GABA_AR-dynamics, as described in ref. 21.

The dynamics of the membrane voltage of excitatory neurons $V_i (i = 1 \dots N_E)$ were given by:

$$C_m \frac{dV_i}{dt} = -g_L(V_i - E_L) - g_{EE,A} \sum_j^{N_E} W_{ij}^{EE} s_j^A (V_i - E_A) - \frac{g_{EE,N}}{1 + e^{-aV_i}/3.57} \sum_j^{N_E} W_{ij}^{EE} s_j^N (V_i - E_N) - g_{IE} \sum_j^{N_I} s_j^G (V_i - E_G) - g_{ext,E} s_{ext} (V_i - E_A) + I_i^s \quad (10)$$

with membrane capacitance $C_m = 0.5$ nF, leak conductance $g_L = 25$ nS, leak reversal potential $E_L = -70$ mV, AMPAR, GABA_AR and NMDAR reversal potentials $E_A = 0$ mV, $E_G = -70$ mV, $E_N = 0$ mV, unitary conductances $g_{ext,E} = 3.1$ nS, $g_{IE} = 2.672$ nS, $g_{EE,N} = 0.56$ nS, $g_{EE,A} = 0.502$ nS, and the NMDAR magnesium block parameter $a = 0.062$ mV⁻¹. In simulations of reduced NMDAR conductance, parameters $g_{EE,N}$ or respectively $g_{EE,A}$ were modulated as indicated in Fig. 3b, c, e, f and Supplementary Fig. 14.

The membrane voltage of inhibitory neurons followed:

$$C_m \frac{dV_i}{dt} = -g_L(V_i - E_L) - g_{EI,A} \sum_j^{N_E} s_j^A (V_i - E_A) - \frac{g_{EI,N}}{1 + e^{-aV_i}/3.57} \sum_j^{N_E} s_j^N (V_i - E_N) - g_{II} \sum_j^{N_I} s_j^G (V_j - E_G) - g_{ext,I} s_{ext} (V_i - E_A) \quad (11)$$

with $C_m = 0.2$ nF, $g_L = 20$ nS, $g_{ext,I} = 2.38$ nS, $g_{II} = 2.048$ nS, $g_{EI,A} = 0.384$ nS and $g_{EI,N} = 0.424$ nS.

The kinetics of synaptic variables $s_i^A (i = 1 \dots N_E)$, $s_i^G (i = 1 \dots N_I)$, and s_{ext} were determined by

$$\frac{ds_x}{dt} = -\frac{s_x}{\tau_x} + w \sum_i \delta(t - t_i) \quad (12)$$

with $\tau_A = 2$ ms, $\tau_G = 10$ ms, $\tau_{ext} = 2$ ms, and the summation running over all spike times t_i so that at each spike time the synaptic variable increased by a step of magnitude w , which was generally set to 1 except for synapses undergoing synaptic potentiation (see below). For s_{ext} , spike times were generated as a Poisson spike train of rate 1800 spikes s⁻¹ (simulating inputs from 1000 external Poisson neurons firing at 1.8 spikes s⁻¹ each).

The slower and saturating NMDAR synaptic variables $s_i^N (i = 1 \dots N_E)$ followed the coupled equations:

$$\frac{ds_i^N}{dt} = -\frac{s_i^N}{\tau_{N_i}} + \alpha_N x_i (1 - s_i^N) \quad (13)$$

$$\frac{dx_i}{dt} = -\frac{x_i}{\tau_{N_x}} + w \sum_j \delta(t - t_j) \quad (14)$$

with $\tau_{N_i} = 100$ ms, $\tau_{N_x} = 2$ ms, and $\alpha_N = 0.5$ kHz.

The strength of recurrent excitatory synapses was modulated depending on the distance in preferred location of presynaptic and postsynaptic excitatory neurons: $W_{ij}^{EE} = J(\theta_i - \theta_j)$, where J is a Gaussian function (centered at $\mu = 0$ with $\sigma = 14.4$ degrees) plus a constant, tuned so that $\sum_j J(\theta_i - \theta_j) = N_E$ and $J(0) = 1.63$. As a

result, neurons with similar preferred locations had 1.63 stronger weights than the average weight (Supplementary Fig. 10 for network scheme and weight profiles).

STP rule in neural network simulations. For connections between excitatory neurons, the spike-triggered step in AMPAR and NMDAR synaptic variables w could vary individually for each specific connection: w_{ij} characterized the step at the synapse from neuron j onto neuron i . Upon synchronized pre- and post-synaptic spiking, w_{ij} was slightly enhanced by an amount Δ_w that depended on the relative spike times of neuron j and i (Fig. 2c) to simulate an increase in probability of glutamate release⁷¹:

$$w_{ij} = w_{ij} + \Delta_w (t_j - t_i) \geq 1 \quad (15)$$

The associative nature of this rule was determined by a potentiation function that required synchronization within a specific temporal window (Fig. 2d):

$$\Delta_w (t_j - t_i) = P \exp(-|t_j - t_i|/\tau_\Delta), \quad (16)$$

with potentiation factor $P = 0.00022$ and $\tau_\Delta = 20$ ms. Changes were sustained (did not decay with time), but synapses depotentiated based on presynaptic activity³: at

each presynaptic spike

$$w_{ij} = w_{ij} - 0.04 * (w_{ij} - 1) \quad (17)$$

Trial structure in neural network simulations. We simulated 20,000 pairs of consecutive trials with independent randomized stimulus locations. Network inputs θ_n^s in trial n with stimulus s were slightly transformed to mimic a repulsive baseline bias away from previous stimulus locations, resulting from sensory aftereffects produced in lower-level cortical areas²⁹: $\theta_n^s = \theta_n^s + 1.25 \text{DoG}(\theta_n^d)$, where $\text{DoG}(\theta_n^d)$ is the first-derivative-of-Gaussian function with $\mu = 0$ and $\sigma = 0.8$ radians, and θ_n^d is the distance between previous and current stimulus.

Simulations started with a stimulus presentation at 0° (trial $n - 1$) for 0.25 s. After the input was removed, a delay of 1 s followed. A negative input to the whole network during 0.25 s simulated the response and removed stimulus-associated neural activity. After an ITI of 3 s, a second stimulus (trial n) was delivered at a random location for 0.25 s. The second delay duration was 3 s. To obtain behavioral readouts from the network, we counted each neuron's spikes during three time windows of 0.25 ms: 0–0.25 s after stimulus offset (0 s delay condition), 0.75–1 s (1 s delay), and 2.75–3 s after stimulus offset (3 s delay). The behavioral response was determined as the angular direction of the population vector of spike counts.

Neural network behavioral analysis. We first calculated the percentage of outlier responses and excluded outlier trials from the network's population vector responses (response error >1 radian). Circular standard deviations and serial dependence were then calculated from the network's population vector responses analogous to human error analyses. In Fig. 3a–f, bias strength was measured as the sum of bias term coefficients in the linear model

$$\theta_n^s = \beta_0 + \beta_{1,d} \text{delay}_n - \beta_2 \text{DoG}(\theta_n^d) - \beta_{3,d} \text{delay}_n \text{DoG}(\theta_n^d) + \epsilon_n \quad (18)$$

that fitted errors θ_n^s in trial n from each parameter manipulation (P , g_{EE} , and g_{EI}) separately as a function of delay and $\text{DoG}(\theta_n^d)$ with $\mu = 0$ and $\sigma = 0.6$ radians.

Hyperparameter cross-validation for neural network responses. The value of hyperparameter σ was determined in a cross-validation procedure for the baseline condition with $P = 0.00022$, $g_{EE} = 0.56$ nS, and $g_{EI} = 0.424$ nS, for values $\sigma \in [0.2, 1.8]$ (in radians). For each value of σ , we fitted the linear model described in Eq. (18) to 67% of trials and tested the prediction in the left-out 33% of trials. Performance for each σ was evaluated using the mean squared error (MSE) of predictions from 1000 cross-validation repetitions. σ was chosen to minimize the MSE of the linear model, yielding $\sigma = 0.6$ radians (Supplementary Fig. 13).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Behavioral data analyzed in this article are openly available by accessing the github repository: github.com/comptelab/serialNMDA

Code availability

Custom code used for simulations and data analysis is openly accessible through the github repository: github.com/comptelab/serialNMDA

Received: 9 December 2019; Accepted: 31 July 2020;

Published online: 25 August 2020

References

- Wang, X. J. Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *J. Neurosci.* **19**, 9587–9603 (1999).
- Wang, M. et al. NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. *Neuron* **77**, 736–749 (2013).
- Volianskis, A. et al. Different NMDA receptor subtypes mediate induction of long-term potentiation and two forms of short-term potentiation at CA1 synapses in rat hippocampus in vitro. *J. Physiol.* **591**, 955–972 (2013).
- Erickson, M. A., Maramba, L. A. & Lisman, J. A single brief burst induces GluR1-dependent associative short-term potentiation: a potential mechanism for short-term memory. *J. Cogn. Neurosci.* **22**, 2530–2540 (2010).
- Castro-Alamancos, M. A. & Connors, B. W. Short-term synaptic enhancement and long-term potentiation in neocortex. *Proc. Natl. Acad. Sci.* **93**, 1335–1339 (1996).

6. Olney, J. W., Newcomer, J. W. & Farber, N. B. NMDA receptor hypofunction model of schizophrenia. *J. Psychiatr. Res.* **33**, 523–533 (1999).
7. Morgan, C. J. A. & Curran, H. V. Acute and chronic effects of ketamine upon human memory: a review. *Psychopharmacology* **188**, 408–424 (2006).
8. Gilmour, G. et al. NMDA receptors, cognition and schizophrenia—testing the validity of the NMDA receptor hypofunction hypothesis. *Neuropharmacology* **62**, 1401–1412 (2012).
9. Catts, V. S., Lai, Y. L., Weickert, C. S., Weickert, T. W. & Catts, S. V. A quantitative review of the postmortem evidence for decreased cortical N-methyl-D-aspartate receptor expression levels in schizophrenia: How can we link molecular abnormalities to mismatch negativity deficits? *Biol. Psychol.* **116**, 57–67 (2016).
10. Hughes, E. G. et al. Cellular and synaptic mechanisms of anti-NMDA receptor encephalitis. *J. Neurosci.* **30**, 5866–5875 (2010).
11. Dalmau, J. et al. Paraneoplastic anti-N-methyl-D-aspartate receptor encephalitis associated with ovarian teratoma. *Ann. Neurol.* **61**, 25–36 (2007).
12. Finke, C. et al. Cognitive deficits following anti-NMDA receptor encephalitis. *J. Neurol. Neurosurg. Psychiatr.* **83**, 195–198 (2012).
13. Maneta, E. & Garcia, G. Psychiatric manifestations of anti-NMDA receptor encephalitis: neurobiological underpinnings and differential diagnostic implications. *Psychosomatics* **55**, 37–44 (2014).
14. Steiner, J. et al. Increased prevalence of diverse N-methyl-D-aspartate glutamate receptor antibodies in patients with an initial diagnosis of schizophrenia: specific relevance of IgG NR1a antibodies for distinction from N-methyl-D-aspartate glutamate receptor encephalitis. *JAMA Psychiatry* **70**, 271–278 (2013).
15. Kayser, M. S. & Dalmau, J. Anti-NMDA receptor encephalitis, autoimmunity, and psychosis. *Schizophr. Res.* **176**, 36–40 (2014).
16. Fischer, J. & Whitney, D. Serial dependence in visual perception. *Nat. Neurosci.* **17**, 738–743 (2014).
17. Kiyonaga, A., Scimeca, J. M., Bliss, D. P. & Whitney, D. Serial dependence across perception, attention, and memory. *Trends Cogn. Sci.* **21**, 493–497 (2017).
18. Barbosa, J. et al. Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nat. Neurosci.* **23**, 1016–1024 (2020).
19. Wimmer, K., Nykamp, D. Q., Constantinidis, C. & Compte, A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neurosci.* **17**, 431–439 (2014).
20. Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* **61**, 331–349 (1989).
21. Compte, A., Brunel, N., Goldman-Rakic, P. S. & Wang, X. J. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb. Cortex* **10**, 910–923 (2000).
22. Kilpatrick, Z. P. Synaptic mechanisms of interference in working memory. *Sci. Rep.* **8**, 7879 (2018).
23. Nakazawa, K., Jeevakumar, V. & Nakao, K. Spatial and temporal boundaries of NMDA receptor hypofunction leading to schizophrenia. *NPJ Schizophr.* **3**, 7 (2017).
24. Yang, G. J. et al. Functional hierarchy underlies preferential connectivity disturbances in schizophrenia. *Proc. Natl Acad. Sci. USA* **113**, E219–E228 (2016).
25. Crabtree, G. W. & Gogos, J. A. Synaptic plasticity, neural circuits, and the emerging role of altered short-term information processing in schizophrenia. *Front. Synaptic Neurosci.* **6**, 28 (2014).
26. Lee, J. & Park, S. Working memory impairments in schizophrenia: a meta-analysis. *J. Abnorm. Psychol.* **114**, 599–611 (2005).
27. Starc, M. et al. Schizophrenia is associated with a pattern of spatial working memory deficits consistent with cortical disinhibition. *Schizophr. Res.* **181**, 107–116 (2017).
28. Fritsche, M., Mostert, P. & de Lange, F. P. Opposite effects of recent history on perception and decision. *Curr. Biol.* **27**, 590–595 (2017).
29. Bliss, D. P., Sun, J. J. & D'Esposito, M. Serial dependence is absent at the time of perception but increases in visual working memory. *Sci. Rep.* **7**, 14739 (2017).
30. Fornaciai, M. & Park, J. Spontaneous repulsive adaptation in the absence of attractive serial dependence. *J. Vis.* **19**, 21 (2019).
31. Bliss, D. P. & D'Esposito, M. Synaptic augmentation in a cortical circuit model reproduces serial dependence in visual working memory. *PLoS ONE* **12**, e0188927 (2017).
32. Driesen, N. R. et al. Impairment of working memory maintenance and response in schizophrenia: functional magnetic resonance imaging evidence. *Biol. Psychiatry* **64**, 1026–1034 (2008).
33. Driesen, N. R. et al. The impact of NMDA receptor blockade on human working memory-related prefrontal function and connectivity. *Neuropsychopharmacology* **38**, 2613–2622 (2013).
34. Kullmann, D. M. & Lamsa, K. P. Long-term synaptic plasticity in hippocampal interneurons. *Nat. Rev. Neurosci.* **8**, 687–699 (2007).
35. Arguello, P. A. & Gogos, J. A. Genetic and cognitive windows into circuit mechanisms of psychiatric disease. *Trends Neurosci.* **35**, 3–13 (2012).
36. Dalmau, J. et al. Anti-NMDA-receptor encephalitis: case series and analysis of the effects of antibodies. *Lancet Neurol.* **7**, 1091–1098 (2008).
37. Gleichman, A. J., Spruce, L. A., Dalmau, J., Seeholzer, S. H. & Lynch, D. R. Anti-NMDA receptor encephalitis antibody binding is dependent on amino acid identity of a small region within the GluN1 amino terminal domain. *J. Neurosci.* **32**, 11082–11094 (2012).
38. Akrami, A., Kopec, C. D., Diamond, M. E. & Brody, C. D. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature* **554**, 368–372 (2018).
39. Masdeu, J. C., Dalmau, J. & Berman, K. F. NMDA receptor internalization by autoantibodies: a reversible mechanism underlying psychosis? *Trends Neurosci.* **39**, 300–310 (2016).
40. Cano-Colino, M. & Compte, A. A computational model for spatial working memory deficits in schizophrenia. *Pharmacopsychiatry* **45** Suppl 1, S49–S56 (2012).
41. Murray, J. D. et al. Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. *Cereb. Cortex* **24**, 859–872 (2014).
42. Jardri, R., Duvernoy, S., Litvinova, A. S. & Denève, S. Experimental evidence for circular inference in schizophrenia. *Nat. Commun.* **8**, 14218 (2017).
43. Itskov, V., Hansel, D. & Tsodyks, M. Short-Term Facilitation may Stabilize Parametric Working Memory Trace. *Front. Comput. Neurosci.* **5**, 40 (2011).
44. Seeholzer, A., Deger, M. & Gerstner, W. Stability of working memory in continuous attractor networks under the control of short-term plasticity. *PLoS Comput. Biol.* **15**, e1006928 (2019).
45. Matute, C. et al. N-methyl-D-aspartate receptor antibodies in autoimmune encephalopathy alter oligodendrocyte function. *Ann. Neurol.* **87**, 670–676 (2020).
46. Konrad, A. & Winterer, G. Disturbed structural connectivity in schizophrenia primary factor in pathology or epiphenomenon? *Schizophr. Bull.* **34**, 72–92 (2008).
47. Finke, C. et al. Functional and structural brain changes in anti-N-methyl-D-aspartate receptor encephalitis. *Ann. Neurol.* **74**, 284–296 (2013).
48. Carter, E. & Wang, X.-J. Cannabinoid-mediated disinhibition and working memory: dynamical interplay of multiple feedback mechanisms in a continuous attractor model of prefrontal cortex. *Cereb. Cortex* **17** Suppl 1, i16–i26 (2007).
49. Kaller, C. P. et al. Working memory in schizophrenia: behavioral and neural evidence for reduced susceptibility to item-specific proactive interference. *Biol. Psychiatry* **76**, 486–494 (2014).
50. Hahn, B. et al. Control of working memory content in schizophrenia. *Schizophr. Res.* **134**, 70–75 (2012).
51. Lieder, I. et al. Perceptual bias reveals slow-updating in autism and fast-forgetting in dyslexia. *Nat. Neurosci.* **22**, 256–264 (2019).
52. Lee, E.-J., Choi, S. Y. & Kim, E. NMDA receptor dysfunction in autism spectrum disorders. *Curr. Opin. Pharmacol.* **20**, 8–13 (2015).
53. Hemsley, D. R. The development of a cognitive model of schizophrenia: placing it in context. *Neurosci. Biobehav. Rev.* **29**, 977–988 (2005).
54. Fletcher, P. C. & Frith, C. D. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat. Rev. Neurosci.* **10**, 48–58 (2009).
55. Corlett, P. R., Honey, G. D., Krystal, J. H. & Fletcher, P. C. Glutamatergic model psychoses: prediction error, learning, and inference. *Neuropsychopharmacology* **36**, 294–315 (2011).
56. First, M. B., Spitzer, R. L., Gibbon, M. & Williams, J. B. W. *Structured Clinical Interview for DSM-IV Axis I Disorders, Clinician Version (SCID-CV)*. (American Psychiatric Press, Inc., 1996).
57. Graus, F. et al. A clinical approach to diagnosis of autoimmune encephalitis. *Lancet Neurol.* **15**, 391–404 (2016).
58. Titulaer, M. J. et al. Treatment and prognostic factors for long-term outcome in patients with anti-NMDA receptor encephalitis: an observational cohort study. *Lancet Neurol.* **12**, 157–165 (2013).
59. Kay, S. R., Fiszbein, A., Vital-Herne, M. & Fuentes, L. S. The positive and negative syndrome scale—Spanish adaptation. *J. Nerv. Ment. Dis.* **178**, 510–517 (1990).
60. Colom, F. et al. [Spanish version of a scale for the assessment of mania: validity and reliability of the Young Mania Rating Scale]. *Med Clin. (Barc.)* **119**, 366–371 (2002).
61. Ramos-Brieva, J. A. & Cordero-Villafafila, A. A new validation of the Hamilton Rating Scale for Depression. *J. Psychiatr. Res.* **22**, 21–28 (1988).

62. Bobes, J., Portilla, M. P. G., Bascarán, M. T., Saiz, P. A. & Bousoño, M. *Banco de instrumentos básicos para la práctica de la psiquiatría clínica*. (Psiquiatría Editores S.L., 2004).
63. Woods, S. W. Chlorpromazine equivalent doses for the newer atypical antipsychotics. *J. Clin. Psychiatry* **64**, 663–667 (2003).
64. Gautier, Laurent. *rpy2*. <https://rpy2.github.io/> (2019).
65. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
66. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* <https://www.jstatsoft.org/article/view/v082i13> (2017).
67. Nash, J. C. & Varadhan, R. Unifying optimization algorithms to aid software system users: optimx for R. *J. Stat. Softw.* <https://www.jstatsoft.org/article/view/v043i09> (2011).
68. Lumley, T., Diehr, P., Emerson, S. & Chen, L. The importance of the normality assumption in large public health data sets. *Annu. Rev. Public Health* **23**, 151–169 (2002).
69. Fletcher, T. D. *Psychometric: Applied Psychometric Theory*. (2010).
70. Stimberg, M., Brette, R. & Goodman, D. F. Brian 2, an intuitive and efficient neural simulator. *elife* **8**, e47314 (2019).
71. Volianskis, A. et al. Long-term potentiation and the role of N-methyl-D-aspartate receptors. *Brain Res.* **1621**, 5–16 (2015).

Acknowledgements

We acknowledge support from Institute Carlos III, Spain (Ref: PIE 16/00014), CELLEX Foundation, Safra Foundation, CERCA Programme/Generalitat de Catalunya, Generalitat de Catalunya (AGAUR 2014SGR1265, 2017SGR01565), “la Caixa” Foundation (ID 100010434, under the agreement LCF/PR/HR17/52150001), and by the Spanish Ministry of Science, Competitiveness and Universities co-funded by the European Regional Development Fund (Refs: BFU 2015-65318-R, RTI2018-094190-B-I00). HS was supported by the “la Caixa” Banking Foundation (Ref: LCF/BQ/IN17/11620008), and the European Union’s Horizon 2020 Marie Skłodowska-Curie grant (Ref: 713673). JB was supported by the Bial Foundation (ref: 356/19). We thank the Barcelona Supercomputing Center (BSC) for providing computing resources. This work was developed at the buildings Centro Esther Koplowitz, CELLEX, and Hospital Clinic, Barcelona. We thank Thaïs Armangué, Domingo Escudero, Amaia Muñoz-Lopetegui, and Gisela Sugranyes for assistance in recruiting patients. Jaime de la Rocha, Daniel Linares for discussions, Ainhoa Hermoso-Mendizabal for comments on the manuscript, and Diego Lozano-Soldevilla for assistance during the development of the task.

Author contributions

H.S., J.B., and A.C. designed behavioral and computational aspects of the study. J.C.F., J.D., and A.C. designed clinical aspects of the study. H.S. performed analyses of human

behavior and computer simulations. H.S., J.B., and A.C. developed the computational model. H.S., A.M., L.P., and A.G.G. performed human experiments. M.R.J. and L.P. performed neuropsychological testing. J.C.F., M.R.J., H.A., and E.M.H. recruited participants for the study. H.S. and A.C. wrote the paper. H.S., J.B., J.D., and A.C. discussed the results and edited the paper. All authors reviewed the paper for intellectual content.

Competing interests

J.D. receives royalties from Athena Diagnostics for the use of Ma2 as an autoantibody test and from Euroimmun for the use of NMDAR, GABA_B receptor, GABA_A receptor, DPPX and IgLON5 as autoantibody tests. All other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-18033-3>.

Correspondence and requests for materials should be addressed to A.C.

Peer review information *Nature Communications* thanks Jardri Renaud and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

3.3 Working memory codes are reactivated between trials in healthy controls, but not in anti-NMDAR encephalitis and schizophrenia

In this section¹⁶, I use EEG decoding techniques to test whether the neural mechanisms described in Chapter 3.1 are altered in anti-NMDAR encephalitis and schizophrenia, given the finding of reduced serial dependence and the hypothesis of reduced short-term potentiation¹⁷ in these patients, as described in Chapter 3.2.

Supplementary material for this Chapter is included in Appendix A3.

¹⁶ This section is a manuscript in preparation. A similar list of authors contributed to this work as in Chapter 3.2. I would like to additionally acknowledge the contribution of Joan Santamaria, who supported the experimental EEG setup in the hospital, and Diego Lozano-Soldevilla, who set up the EEG preprocessing pipeline for this project.

¹⁷ Note that in Chapters 3.2 and 3.3, I will use the abbreviation STP to refer to NMDAR-dependent short-term potentiation, in contrast with the remainder of this thesis, in which STP is used to refer to short-term plasticity.

Working memory codes are reactivated between trials in healthy controls, but not in anti-NMDAR encephalitis and schizophrenia

Introduction

Working memory is a brain function that lies at the core of cognition ¹, as it maintains information for short periods of time in a format that makes it easy to access and manipulate ². Working memory deficits occur across a wide range of brain disorders, including major depression, bipolar disorder, ADHD, and Parkinson's disease ³. Yet, they might play a particularly central role for thought disorders in specific psychiatric and neurological diseases, such as schizophrenia ⁴ or anti-NMDA-receptor (NMDAR) encephalitis ⁵, where working memory deficits even form part of early diagnostic criteria ⁶. Memory alterations in both anti-NMDAR encephalitis and schizophrenia could result from an underlying dysfunction of the NMDAR in memory-maintaining circuits like the prefrontal cortex ⁷⁻⁹. This hypothesis is supported by working memory studies under pharmacological blockade of the NMDAR ^{10,11}, computational modeling of NMDAR dysfunction in memory-maintaining cortical circuits ^{12,13}, post-mortem assessments of NMDAR densities in prefrontal cortex in patients with schizophrenia ¹⁴, and the drastic reduction of NMDAR levels caused by anti-NMDAR encephalitis ¹⁵. In line with this hypothesis, we have recently reported a novel, qualitative alteration in working memory processing in patients with anti-NMDAR encephalitis and schizophrenia ¹⁶. Specifically, we showed that while working memory in healthy controls was positively biased towards previous memories (so-called serial dependence), these biases were drastically reduced in both patient groups.

Serial dependence is a common phenomenon in neurotypical subjects ¹⁷ which demonstrates that memory information is carried over between subsequent working memory trials, even when this is not required to perform the task. Given that after the response, memory circuits such as prefrontal cortex usually cease to encode working memory items in persistent neural firing ¹⁸, alternative mechanisms ^{19,20} or brain regions ²¹ have been proposed to represent previous trial items between subsequent trials. We have recently demonstrated that a plasticity-like mechanism can temporarily enhance neural excitability at previously delay-active prefrontal neurons, and that these traces of enhanced excitability can bias upcoming memories towards previous items ²². We explained these findings through a hybrid model of prefrontal cortex that combines rate-based coding during delay and a short-term plasticity mechanism ^{23,24} that spans inter-trial intervals (ITIs) and biases incoming memories.

An important feature of models that combine persistent firing and plasticity in the same circuit is the possibility to retrieve former attractors after the network has returned to a state of spontaneous firing ^{25,26}. When a short, sufficiently strong input is delivered to a random subset of neurons in the network, previously active neurons can return to a pattern of persistent firing. In our recent analyses of single-unit data from monkey prefrontal cortex, we showed that such code reactivations can occur between trials, just before the onset of a new stimulus ²². In a parallel human experiment, we demonstrated that reactivations can

also be measured in alpha power in human electroencephalography (EEG). Memory reactivations in monkeys and humans were triggered by internal or external attention signals, such as ramping neural activity in monkey prefrontal cortex neurons, and the onset of the next trial's fixation period in healthy controls. Importantly, the strength of code reactivations predicted how strongly upcoming memories were biased towards previous ones in both monkeys and humans.

Here, we aimed at testing whether previous-trial working memory codes in multivariate EEG alpha power would be reactivated in the inter-trial interval (ITI) in patients with anti-NMDAR encephalitis and patients with schizophrenia, similar to the reactivations observed in healthy controls in ref. ²². This question was based on two observations: First, we have reported drastically reduced serial dependence in patients with anti-NMDAR encephalitis and schizophrenia ¹⁶, and a link between memory reactivations in EEG and the strength of serial dependence ²². Second, we have postulated that short-term potentiation (STP) mechanisms in memory-maintaining circuits could be disturbed in both patient groups, potentially as a result of NMDAR dysfunction ^{27,28}, leading to a disruption of information carry-over between consecutive trials ¹⁶. Here, we first formalized our hypothesis in a circuit model of prefrontal cortex with varying degrees of STP ¹⁶, where a deficit in STP is reflected in reduced memory reactivations during the ITI. We then analyzed working memory codes in multivariate EEG alpha power, and measured reduced and less stable memory codes during the delay in both patient groups, when compared to healthy controls. We then tested our prediction of reduced memory reactivations during ITI, and found that memory reactivations occurred in healthy controls, but were indeed disrupted in patients with anti-NMDAR encephalitis or schizophrenia. In healthy controls, the strength of memory reactivations correlated with their predictiveness of serial dependence in the upcoming trial on a trial-by-trial basis, while this was not the case in patients, who lacked memory reactivations in the first place. Together, these findings suggest that a lack in between-trial memory reactivations contributes to a reduction in serial dependence, and results from dysfunctional STP mechanisms in patients with anti-NMDAR encephalitis and schizophrenia.

Results

Similar working memory precision, but decreased serial dependence in patients

To measure the neural correlates of working memory maintenance and serial dependence in healthy controls and patients, we conducted a visuospatial working memory experiment in a sample of $n = 22$ controls, $n = 27$ patients with anti-NMDAR encephalitis, and $n = 19$ patients with schizophrenia. Patients with encephalitis had received immunotherapy and were in a prolonged recovery phase, while patients with schizophrenia were tested during a stabilized period. In two sessions of 1.5 h each, subjects had to remember randomized angular locations at fixed eccentricity presented on a computer screen for short delay periods of 0, 1, or 3 s (Methods, Fig. 1a). After subjects indicated the memorized location in one trial with a mouse click, a fixation period of 1.1 s started, before a new, uncorrelated location was presented in the next trial. Here, we recorded EEG activity during delay, and

response and fixation periods at 43 scalp electrodes and analyzed single-trial behavioral and EEG data.

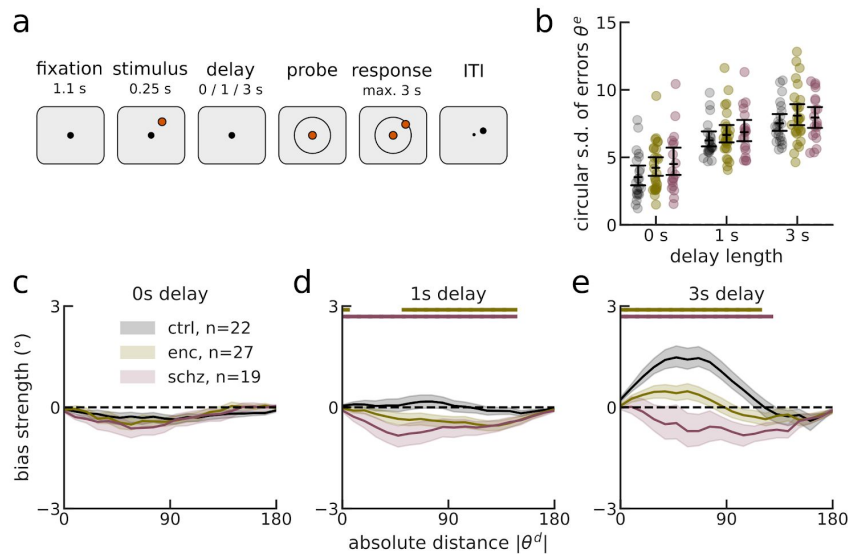


Fig. 1 | Similar working memory precision, but decreased serial dependence in patients

a In each trial, participants were asked to maintain eye fixation and the mouse pointer on a central fixation dot. After 1.1 s (fixation period), a stimulus appeared for 0.25 s at a randomly chosen angular location with fixed distance from the center. Delay lengths varied randomly between trials (0, 1 or 3 s), where 1 s trials occurred with higher probability ($P = 0.67$), and 0 and 3 s trials with equal probabilities ($P = 0.17$). Subjects reported the remembered location with a mouse click and started the next trial by moving the mouse back to the fixation dot during the inter-trial-interval (ITI). **b** Precision for each subject and delay was inversely estimated as the circular s.d. of bias-corrected error distributions (Methods). For longer delays, participants' responses were less precise (*delay*, $F(2,195) = 78.61$, $p < 1e-16$). There were no overall or delay-dependent group differences in precision (*group*, $F(2,195) = 2.62$, $p = 0.75$; *group* \times *delay*, $F(4,195) = 0.16$, $p = 0.96$, all p-values from ANOVA). Error bars indicate 95% bootstrap C.I. of the mean. **c-e**, Serial dependence by group and delay length. Serial dependence is calculated as the 'folded' error $\theta^{e'}$ for different θ^d (Methods). Shading, 32% bootstrap C.I. of the mean across participants for $n = 22$ healthy controls (ctrl), $n = 19$ patients with schizophrenia (schz), and $n = 27$ patients with anti-NMDAR encephalitis (enc). **c**, Serial dependence for all groups was repulsive in 0 s trials. **d** For 1 s trials, group differences in serial dependence emerged between ctrl and enc, and between ctrl and schz. **e** After 3 s delay, both patient groups showed drastically reduced biases compared to ctrl. Significance bars show significant permutation tests (1,000 permutations) at $\alpha = 0.05$ between groups.

After analyzing behavioral measures of working memory in ¹⁶, we included further subjects to increase the sample for the EEG study. We therefore asked whether previously reported behavioral results could be confirmed in the extended sample. First, we measured the circular standard deviation of responses around the target for each subject and each delay. This measure is an inverse estimate of memory precision and increases with delay length, as a function of accumulating noise in working memory delay activity ²⁹. We used a linear model to compare the delay-dependent decrease in precision between groups (Fig. 1b). Similar to our previous findings, precision in patients did not decrease more strongly with

delay, and there were no overall differences in precision between groups (*group*, $F(2,195) = 2.62$, $p = 0.75$; *group* \times *delay*, $F(4,195) = 0.16$, $p = 0.96$, ANOVA).

We also calculated serial dependence in our extended sample to confirm previous results. We averaged errors around the target separately for different, binned distances between previous and current targets (Methods). Positive mean folded errors in these curves denote attraction to previous targets, while negative values denote repulsion. Fig. 1c-e shows mean “folded” bias curves for each group and delay. In congruence with our previous results¹⁶, we found small, repulsive biases for all groups in conditions with low memory requirements (0 s delay), and gradually increasing attractive biases in healthy controls for longer working memory delays. In contrast, biases remained repulsive for all delay lengths in patients with schizophrenia, and became only minimally attractive in encephalitis patients for 3 s delays.

STP moderates reactivation strength in a circuit model of working memory

In our previous work, we simulated reduced serial dependence in patients as an effect of reduced NMDAR-dependent STP¹⁶. However, local STP-like mechanisms or the absence thereof cannot be measured in scalp EEG, which limits our ability to test this hypothesis in patients. Therefore, here we used the model proposed in ref.¹⁶ to derive predictions about the neural code that should be observed if STP was disrupted in patients. In particular, we were interested in working memory reactivations occurring between trials: In ref.²², we have demonstrated that circuits that combine persistent delay activity and STP-based, “silent” memory traces can produce reactivations of memory codes when a non-specific input is delivered to all neurons, and showed that such reactivations occur in monkey prefrontal cortex neurons, and in EEG alpha power in humans.

We then tested whether memory code reactivations were indeed disrupted in the circuit model proposed in ref.¹⁶ when STP levels were modulated to mimic the behavioral alterations observed in our patient populations (Fig. 2a,b). In this model, working memory representations are held in persistent delay firing (Fig. 2c), and delay activity leaves potentiated synaptic traces (Fig. 2d) that later bias upcoming delay activity towards old memories. We hypothesized that NMDAR dysfunction in patients is reflected in reduced STP (Fig. 2b,d), while synaptic conductances through NMDARs remain intact. To simulate memory reactivations in the inter-trial interval (ITI), we delivered a non-specific, excitatory drive to all neurons of the network (orange triangle in Fig. 2c,d). This drive might reflect external signals, such as the onset of the fixation dot, or internal preparatory signals, such as attention processes. In line with our previous results²², memory codes were reactivated in some trials, leading to an increase in firing rate tuning during the fixation period, before the onset of the next stimulus (black line in Fig. 2c). In contrast, when STP was reduced (green and lilac dashed lines in Fig. 2d), memory reactivations occurred less frequently (Fig. 2c). Hence, our modeling predicts that memory codes should be relatively intact during delay; however, memory reactivations in the ITI should be significantly reduced in patients with anti-NMDAR encephalitis and patients with schizophrenia, if STP was disrupted in these diseases.

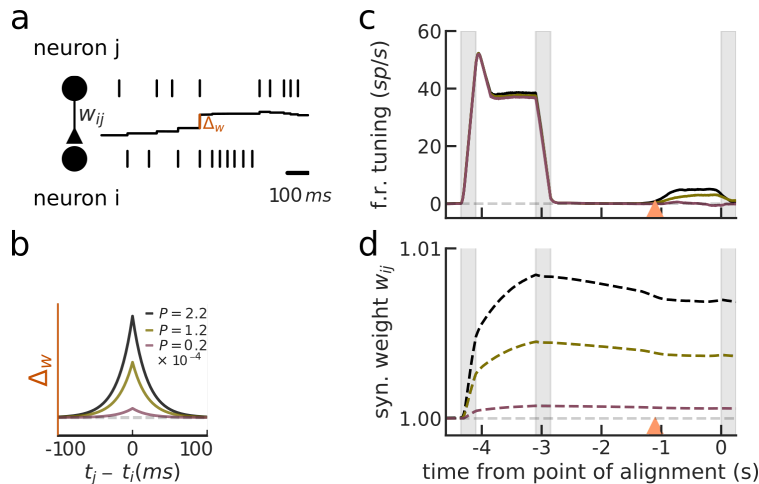


Fig. 2 | Reduced STP in a circuit model of working memory disrupts memory reactivations during ITI

To test the impact of STP on memory code reactivations, we simulated consecutive working memory trials in a network model that combines persistent delay activity and STP. **a** STP was implemented as an activity-dependent increase and decrease in synaptic weights between excitatory model neurons. The strength of each individual synapse is determined by w_{ij} (a, middle black trace), which is potentiated at each spike by an amount Δ_w that depends on the relative spike times t_j and t_i of pre- and postsynaptic neurons, respectively, and on the potentiation factor P and it is reduced by an amount relative to the synaptic strength at each presynaptic spike, resulting in activity-dependent decay. **b** Different potentiation factors P are chosen to represent reduced STP in anti-NMDAR encephalitis and schizophrenia. **c** Firing rate tuning (calculated as the difference in average firing rates in 21 neurons centered at the presented stimulus value and 21 neurons centered at the opposite stimulus value) and **d** weight traces (averaged over 21 stimulus-selective neurons) in simulations with high (black), gradually reduced (green) and drastically reduced (lilac) values of potentiation factor P , as shown in b. Firing rate tuning and weights traces are averaged over 1,000 trials. When stimuli in trial $n-1$ were presented at the neurons' preferred location, a subgroup of neurons represented the memory in a stable working memory code, discernible in firing rate tuning. Through the increase in synchronized firing, weights w were potentiated between delay-coactive neurons. After firing rates returned to spontaneous levels, firing rate tuning was lost, but synaptic weights remained potentiated until the onset of the next stimulus. When a short, unspecific drive was delivered to all neurons in the network at -1.1 s, memory codes from the previous trial were recovered into neural tuning in some trials for networks with high- and intermediate STP levels, but not in networks with disrupted STP.

Working memory contents during delay are decoded with less reliability

Before testing our hypothesis regarding neural reactivations during ITI, we characterized working memory codes during the delay period in all three groups. To this end, we extracted alpha power (8-12 Hz) from EEG signals at 43 scalp electrodes. Alpha power decreases in occipital sites contralaterally to attended locations and for locations being actively maintained in working memory^{30,31}. We used this characteristic to decode memorized spatial locations from the spatial distribution of alpha power across electrodes. We assigned stimuli to one of eight bins, and trained a linear decoder for each subject on all but one trial. We then used the left-out trial to test the predictive performance of the

decoder, measured as the cosine of the difference between predicted and presented angle (Fig. 3a, Methods). Decoders were first trained and tested at the same time point, to read out the overall information about the stimulus that was available at each period of the trial (we term this kind of decoder a “diagonal decoder”). In Fig. 3, we combine decoding results for 1 and 3 s trials up to 1.25 s after the stimulus onset. Decoding performance throughout the whole 3 s-delay period in 3 s trials (which occurred with a lower probability, thereby significantly reducing the number of trials that can be used for training and testing) is reported in Supplementary Fig. 1.

Spatial locations could be decoded throughout the 1 s delay in all three groups (Fig. 3b, upper significance bars). Memory codes in 1 s trials were weaker in patients than in healthy controls during most time points (lower significance bars). We wondered whether this group difference might be mediated by group differences in the number of trials (ctrl, $n = 900.63 \pm 103.02$; enc, $n = 870.44 \pm 141.93$; and schz, $n = 833 \pm 153.33$ trials): To test this hypothesis, we compared the decoding performance of decoders during the delay period after including increasing numbers of trials (Supplementary Fig. 2). We found that significant differences in delay decoding between healthy controls and patients persisted across different numbers of trials. Moreover, decoding performance saturated after ~ 400 trials in all groups, further strengthening the independence of group differences from the number of trials. Finally, we asked whether differences in decoding performance might be an effect of increased noise in patients’ EEG signals. To answer this question, we conducted a time-frequency analysis to compare task-induced, univariate changes in alpha power between groups (Supplementary Fig. 3). We did not find a deficit in alpha modulation in patients, compared to healthy controls: In fact, alpha power during the working memory delay decreased more strongly in encephalitis patients than in healthy controls. We thus conclude that reduced decoding performance during the memory trial can not be attributed to external factors, such as noise or a less efficient modulation of EEG alpha power through the task, but rather reflects a significant difference in neural processing of working memory contents in patients with anti-NMDAR encephalitis and schizophrenia.

Delay-, but not stimulus-related code components are weaker and less stable in patients

In Fig. 3b, we have shown that memory codes are reduced in patients, compared to healthy controls. However, decoding performance in patients was comparable during some periods of the delay (e.g, shortly before the response probe), while differences were largest in early and mid-delay. To test which aspects of the memory code are affected in patients, we trained cross-temporal decoders^{22,32,33} to disentangle temporal code components that could be related to different processing stages. For each subject, we trained and tested decoders at all combinations of time points from 0.25 s before stimulus onset to the time of the response probe (1.25 s after the stimulus; Fig. 3d-f; Supplementary Fig. 4 shows cross-temporal decoding matrices for 3 s trials), and from 0.25 s before until 0.25 s after the motor response. For each group, we assessed temporal clusters of significant decoding across subjects (red contours), which differed in two aspects: First, and in line with Fig. 3b, we found group differences in the overall decoding performance, especially during the delay period of the trial. In addition, we observed a reduced temporal stability of memory

codes in both patients groups, as compared to healthy controls: While healthy controls' memory codes generalized across all time points of the delay, memory code generalization in patients was temporally more limited to adjacent time points, when using a decoder trained in "mid-delay" (in a window from 0.65 - 0.85 s after stimulus onset, indicated with an orange mark in Fig. 3b,g, and indicated with a dashed white line that marks the window's center in Fig. 3d-f). Finally, we asked whether a reduced memory component could account for overall reduced decoding performance in patients (Fig. 3b). We measured the "residual" code, after accounting for delay decoding, by calculating subject-wise differences between diagonal (Fig. 3b) and delay decoding performance (Fig. 3g). This residual code, which mostly consisted of a stimulus- and a probe-related component, did not differ between patients and healthy controls (Fig. 3h). Hence, we identified a specific decrease in memory codes, but not visual or motor-related codes during the performance of our visuo-spatial working memory task.

Delay codes do not correlate with working memory precision

In Fig. 1b, we have seen that there are no group differences in working memory precision; yet, memory codes during delay differ significantly in strength and temporal stability. Therefore, we designed a more sensitive analysis on the intraindividual level: Again, we trained a trial-wise decoder on the average signal during mid-delay (0.65 s - 0.85 s). We then divided trials in high-decoding and low-decoding based on delay decoding performance (Methods), and calculated the difference between the averages of absolute errors $|\theta^e|$ for each subject (Fig. 3c). This difference did not deviate from zero for all groups, indicating the lack of a relation between decoding strength during delay and behavioral precision in single trials. The absence of a relationship between delay codes and working memory precision at both the group level and the intra-individual is a striking finding that can potentially be explained by the low cognitive demand posed by our task, as detailed further in the discussion.

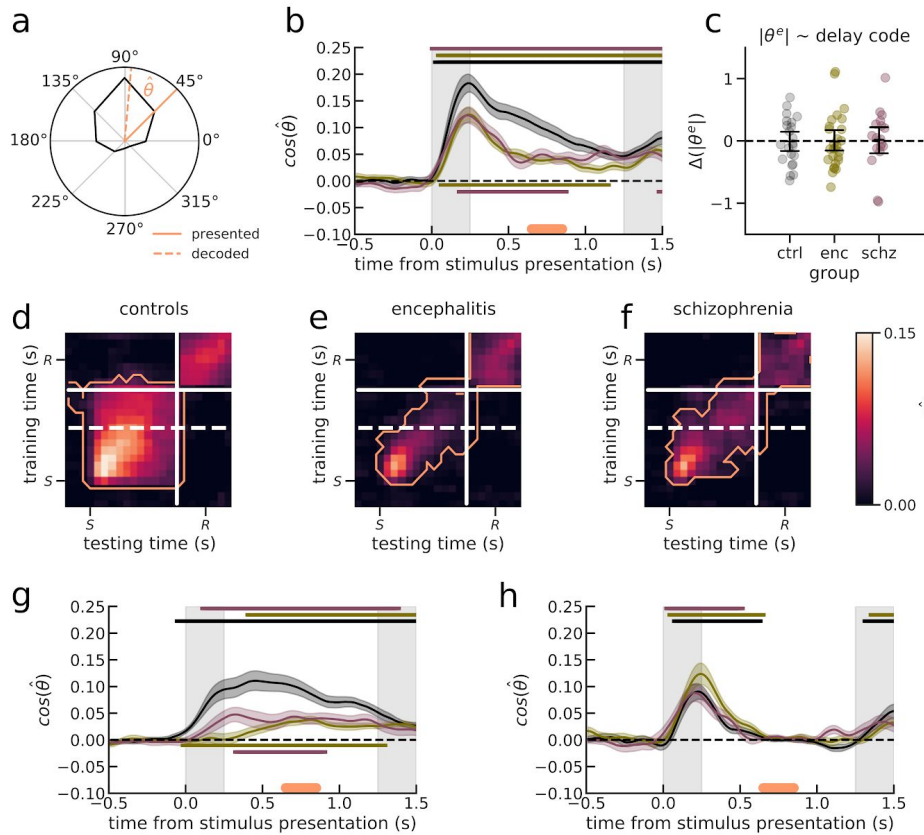


Fig. 3 | Working memory codes in patients are less reliable and less stable during delay

a We calculated decoding performance as the cosine of the angular difference $\hat{\theta}$ between predicted (here, $\sim 85^\circ$) and presented (45°) angles in each trial. The predicted angle was determined as the circular mean of single-trial tuning curves (Methods). **b,g,h** Shading, 32% bootstrap C.I. of the mean across participants for $n = 22$ healthy controls (ctrl), $n = 19$ patients with schizophrenia (schz), and $n = 27$ patients with anti-NMDAR encephalitis (enc); upper significance bars mark significant decoding, lower significance bars mark significant group differences, corrected with 1- or respectively 2-sample cluster permutation tests at $\alpha = 0.05$. Grey bars indicate stimulus and probe presentation, and orange bars indicate the timepoints later referred to as “mid-delay” (0.65-0.85 s). **b** Spatial locations in working memory were decoded throughout the complete 1 s memory delay in all groups. During mid-delay, decoding performance was lower for patients than for healthy controls. **c** Difference in absolute errors as a function of trialwise decoding performance in mid-delay. For each subject, we separated high (>75th percentile) and low-decoding (<75th percentile) trials and tested whether subjects responded with higher precision in high-decoding trials. The difference was non-significant for all groups. **d,e,f** Cross-temporal decoding during stimulus presentation (S), delay, and response (R). White lines mark the discontinuity of EEG signals after the probe onset at 1.25 s. Orange lines mark significant decoding clusters (1-sample permutation test at $\alpha = 0.05$). Spatial locations were decoded above chance at all time points comprising the diagonal for all groups. In controls, memory codes trained at any period of the delay generalized to any other delay time point, while in patients, memory codes did not generalize as extensively. During response, a different code component than during delay represented memory contents in all groups. **g** Temporal generalization of memory codes trained in mid-delay (orange bar). **h** Residual code, when subtracting the delay component (g) from the diagonal decoder (b), does not differ between groups.

Working memory delay code in patients is not reactivated during the ITI

After characterizing working memory codes during delay, we now wanted to test our prediction that group differences in serial dependence, and a disruption in STP, should be reflected by reduced memory code reactivations in patients, compared to healthy controls. To measure reactivations, we used a cross-temporal decoder to read out memory contents throughout two consecutive trials $n-1$ and n . We decoded the location presented in trial $n-1$ throughout the delay period and motor response of trial $n-1$, and the fixation period of trial n , up until the onset of the new stimulus (Fig. 4a-c, where the bottom left part of the matrices corresponds to Fig. 3d-f). Memory contents from the last trial were still decodable in the fixation period, and faded several hundred ms before the onset of the new stimulus. However, the cross-decoder revealed that these memory codes in patients did not share information with the previous delay-period: Rather, memories were represented by a distinct, independent code component (Fig. 4b,c). In contrast, memory reactivations in healthy controls were driven by the previous trial's delay code, reflected in above-chance decoding performance in the lower right part of the decoding matrix in Fig. 4a. Fig. 4d compares a decoder trained in mid-delay (0.65 - 0.85 s after stimulus in trial $n-1$) across groups. These analyses confirm our hypothesis, linking reduced fixation-period memory reactivations (Fig. 4a-d) to disrupted serial dependence (Fig. 1c-e) on the group level.

Strong memory code reactivations predict biases in the upcoming trial

Finally, we asked whether we could relate the strength of memory reactivations to serial dependence on a within-subjects level. We have previously reported such a relationship in a subset of healthy controls²². We thus first assessed this relation in our extended sample of healthy controls: For each subject, we separated trials with high and low reactivation of the delay code during fixation (orange mark in rightmost panel of Fig. 4d), and calculated serial dependence separately for each set of trials (Methods). In Fig. 4e-g, we show the difference Δ_{bias} between high- and low-decoding trials for each subject. Δ_{bias} was positive in healthy controls, but non-significant (Fig. 4e). We wondered whether this relation was modulated by a difference in reactivation strength: In subjects whose memory code does not reemerge, a split based on decoding strength would separate trials mainly based on random fluctuations. We therefore related the difference between high- and low-decoding trials with the strength of memory reactivations for the three different groups. A linear model of Δ_{bias} as a function of *reactivation strength*, *group*, and their interaction confirmed our hypothesis: The higher the average reactivations, the more pronounced were differences between high- and low-decoding trials (main effect of *reactivation strength* $F(62,1) = 9.83$, $p = 0.003$). Although there was no significant interaction between *reactivation strength* and *group* ($F(62,2) = 1.54$, $p = 0.22$), groupwise correlations showed that the main effect of *reactivation strength* was driven mainly by healthy controls, whose reactivations were pronounced (Pearson's $r = 0.60$, $p = 0.003$; Fig. 4e), compared to patients with encephalitis (Pearson's $r = 0.07$, $p = 0.75$) and schizophrenia (Pearson's $r = 0.18$, $p = 0.47$; Fig. 4f,g). To conclude, these results reveal that memory reactivations in EEG alpha power can increase serial dependence in subjects whose memory codes in the fixation period are reliably decoded. Together with the absence of memory reactivations and serial dependence on the group

level, these findings confirm the predictions from our network model with reduced STP in anti-NMDAR encephalitis and schizophrenia.

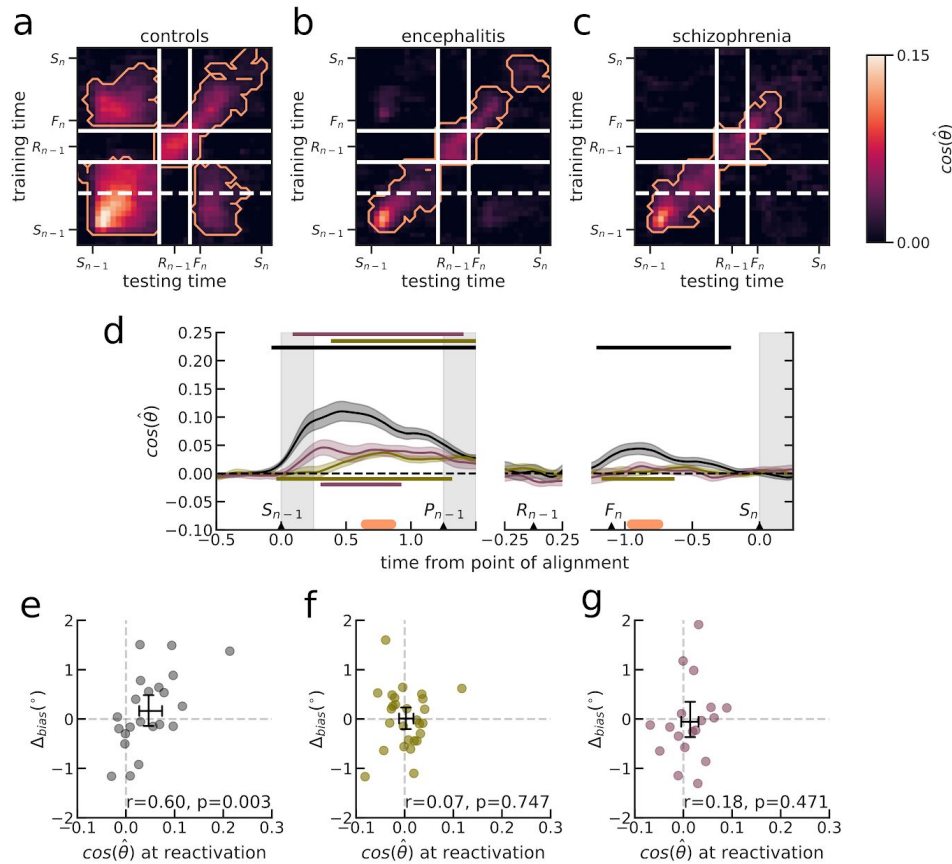


Fig. 4 | Memory codes are reactivated during ITI in healthy controls, but not in patients

a,b,c Cross-temporal decoding over consecutive trials. We decoded the location presented in trial $n-1$ from the time of stimulus presentation (S_{n-1}), over the delay, during the response period (R_{n-1}), and the fixation period in the current trial, starting at fixation dot onset (F_n) and ending at the onset of the new stimulus (S_n). White lines mark the discontinuity of EEG signals after the probe onset at 1.25 s, and between response and fixation periods. Orange lines mark significant decoding clusters (1-sample permutation test t alpha = 0.05). Spatial locations were decoded above chance at all time points comprising the diagonal for all groups, up until S_n in ctrl and enc, and until mid-fixation in schz. Memory codes trained during delay reemerged during fixation in controls, but not in patients, after being interrupted by the response code. **d** Memory codes trained in mid-delay (0.65 - 0.85 s) generalized across large parts of the delay in trial $n-1$ (from stimulus presentation, S_{n-1} , to the response probe P_{n-1}), before disappearing around the time of response (R_{n-1}) and reemerging after the onset of fixation (F_n ; upper significance bars, 1-sample cluster permutation test α = 0.05). Lower significance bars indicate significant group differences between healthy controls and patients (2-sample cluster permutation test α = 0.05). Shading, bootstrap s.e.m. across participants for $n = 22$ healthy controls (ctrl), $n = 19$ patients with schizophrenia (schz), and $n = 27$ patients with anti-NMDAR encephalitis (enc). Grey bars indicate stimulus and probe presentation, and orange bars indicate the timepoints referred to as “mid-delay” and “fixation”. **e** Difference Δ_{bias} in serial dependence between trials with high and low memory reactivations (Methods), as a function of the strength of memory reactivations. Δ_{bias} was positive, but non-significant in healthy controls ($t = 1.00$, $p = 0.33$), and increased with the overall strength of memory reactivations (Pearson’s $r = 0.60$, $p = 0.003$). **f,g** Δ_{bias} was non-significant in both enc ($t = 0.09$, $p = 0.93$) and schz ($t = -0.3$, $p = 0.76$), and did not correlate with overall memory reactivation strength.

Discussion

In this study, we assessed whether reduced serial dependence in anti-NMDAR encephalitis and schizophrenia (Fig. 1) would be reflected in disrupted EEG reactivations of previous-trial memory contents during the inter-trial interval (ITI), as predicted by a model with reduced STP in memory-maintaining circuits (Fig. 2). Moreover, we explored qualities of the working memory code during memory delays in patients, as compared to healthy controls. We showed that multivariate alpha power tracked memory contents throughout short delay periods in all groups. In patients with encephalitis and patients with schizophrenia, memory codes were selectively impaired already in delay, when compared to healthy controls, and generalized less in time (Fig. 3). In all groups, memory codes trained in delay disappeared completely during the response period, to then reappear after the onset of the next trial's fixation dot in healthy controls, but not in patients (Fig. 4). We reasoned that the reactivation of memory codes should be related to serial dependence, and confirmed this prediction in healthy controls: Strong memory code reactivations during the fixation period were related to stronger serial bias in the upcoming trial. In contrast, we did not observe such a relationship in patients with encephalitis or schizophrenia, whose memory codes did not reactivate in the fixation period and were not predictive of biases.

Our findings confirm the relation between working memory reactivations in the ITI and serial dependence, both on the group level and the intra-individual level. Our modeling suggests that both serial dependence and memory reactivations depend on a STP mechanism that is disrupted in patients with anti-NMDAR encephalitis and in patients with schizophrenia, so that both patient groups show reduced serial dependence, and weaker ITI memory reactivations in our working memory task. Following the logic of our network model, memory reactivations are not needed to produce serial dependence: Even in the absence of reactivations, memory traces in STP can still bias upcoming memories¹⁶. Instead, memory reactivations can modulate the strength of biases, as explicitly shown in Fig. 4e, and in ref.²². Hence, in our model, the necessary component for both memory reactivations and serial dependence is a STP-like mechanism. By showing that both serial dependence and memory reactivations are disrupted in patients with anti-NMDAR encephalitis and schizophrenia, we deliver more evidence for the dysfunction of such a STP-like mechanism in the two diseases.

To confirm our hypothesis about how neural computations differ between groups to disrupt memory reactivations during ITI, further experimental work is needed. Combined electrophysiology and pharmacology in animal models could elucidate questions that remain unanswered in this study: Similar to ref.²², the analysis of single-neuron activity during working memory performance could identify neurons that participate in working memory codes, and detect correlations of these neurons' activity between delay and ITI periods. Importantly, in such studies, STP traces between delay-active neurons could be estimated during ITI. The pharmacological blockade of NMDAR could then test hypotheses regarding both STP and the nature of reactivated memory codes that could underlie the observed phenomena in anti-NMDAR encephalitis and schizophrenia. Experimental work would also be needed to discard an alternative explanation of our findings: Working

memory codes could be held in one area during the delay, and transferred to another area after the response. From that area, stimulus-specific activity could be fed back to the memory circuit after the ITI, to then actively influence new, upcoming memories. This interpretation would be congruent with findings of ITI-, but not delay-representations in PPC in ref. ²¹. In this view, a failure to reactivate memories in the fixation period could reflect the disrupted bi-directional communication between memory-maintaining circuits, and brain areas that represent past memory contents during response or inter-trial periods.

Importantly, we cannot offer a straightforward way to interpret reduced and less stable delay codes in patients, relative to healthy controls. With respect to the lower decoding performance during delay, we have shown that our findings are not explained by a reduced number of trials or a reduction in univariate alpha power (Supplementary Figs. 2,3). In fact, alpha power during delay was modulated more strongly in patients with encephalitis than in healthy controls. Moreover, we failed to find a relationship between delay code and working memory precision both on the inter-group level, and on the intra-individual level (between trials). In ref. ¹⁶, we have argued that intact memory precision in patients might be explained through ceiling effects that emerge in the very simple working memory task performed in this study: Subjects only needed to remember one item, during rather limited delay periods, while reduced working memory precision has been found in tasks with long delays ³⁴ or more challenging stimuli ³⁵. In line with this reasoning, ref. ³⁶ have reported reduced memory-related BOLD signals in PFC and reduced memory precision in patients with schizophrenia in a task with long working memory delays. Thus, our task might not have tested the upper limits of working memory. Consequently, variability in response precision or in working memory decoding performance between trials or between individuals might not reflect a cognitively relevant dimension.

To conclude, we identified a disruption in memory code reactivations during ITI in patients with anti-NMDAR encephalitis and schizophrenia. We argued that disruptions in memory reactivations and reduced serial dependence observed in these patients are both related to a disruption in STP. Moreover, memory reactivations and their absence in anti-NMDAR encephalitis and schizophrenia could also directly impact the (lacking) generation of serial dependence, a hypothesis based on observations in previous work in healthy human subjects and monkeys ²², and supported by computational modeling of memory-maintaining circuits ^{16,22}. To understand whether and how the absence of memory reactivations is related to STP and to altered delay period memory codes, further work, ideally accompanied by electrophysiological recordings in prefrontal cortex, is needed that specifically tests this hypothesis.

Methods

Experimental sample

We included $n = 27$ patients with anti-NMDAR encephalitis (enc; age 28.7 ± 11.3 years, mean \pm s.d.; $n = 5$ male), $n = 19$ patients with schizophrenia or schizoaffective disorder (schz; age 21.3 ± 8.8 years, mean \pm s.d.; $n = 8$ male), and $n = 22$ neurologically and psychiatrically healthy control participants (ctrl; age 24.9 ± 10.4 years, mean \pm s.d.; $n = 4$ male), all with normal or corrected vision.

Behavioral data from $n = 19$ controls, $n = 16$ patients with encephalitis, and $n = 17$ patients with schizophrenia has been reported before ¹⁶, and behavioral and EEG data from $n = 14$ healthy controls has been included in another previous study ²². Psychiatric diagnoses (or the absence thereof for controls) were confirmed using the Structured Clinical Interview for DSM IV (SCID-I) ³⁷. Patients with anti-NMDAR encephalitis were recruited from centers in Spain ($n = 25$ in Spain), Germany ($n = 1$), and the United Kingdom ($n = 1$) and participated in the experiment several months after hospital discharge. Patients with anti-NMDAR encephalitis were diagnosed by confirmation of CSF IgG antibodies against the GluN1 subunit of the NMDAR ⁶. All healthy controls and patients with schizophrenia tested seronegative for antibodies against NMDAR in serum ³⁸. Anti-NMDAR encephalitis is known to have a prolonged process of recovery after the acute stage of the disease ³⁹, and patients in the prolonged recovery phase still suffer from cognitive deficits as has been previously described in cohorts with long follow-up ⁴⁰. All patients were sufficiently recovered to participate in the testing procedure. Controls and patients with schizophrenia were recruited from the Barcelona area and from Hospital Clínic (Barcelona, Spain), respectively. Patients with schizophrenia were clinically stable at the time of testing. All participants (and, in the case of minors of age, their legal guardians) provided written informed consent and were monetarily compensated for their time and travel expenses, as reviewed and approved by the Research Ethics Committee of Hospital Clínic. All subjects were assessed for psychiatric symptoms and functionality through a battery of standard tests including the Spanish versions of the Positive and Negative Syndrome Scale (PANSS) ⁴¹, the Young Mania Rating Scale (YMRS) ⁴², the Hamilton Depression Rating Scale (HAM-D) ⁴³ and the Global Assessment of Functioning Scale (GAF) ⁴⁴.

Task protocol and behavioral testing

Participants completed two 1.5 h sessions performing a visuospatial working memory task described in Fig. 1a. In each session, participants were asked to complete 12 blocks of 48 trials. However, some participants did not complete all blocks, and some did not complete both sessions (on average, participants completed 1138.5 trials (median, ctrl, i.q.r. = 35.0), 1101.0 trials (median, enc, i.q.r. = 110.0), and 1017.0 trials (median, schz), i.q.r. = 301.0). For stimulus presentation, we used Psychopy v3.1.5 on Python 2.7, running on a 17" HP ProBook laptop. Each trial began with the presentation of a central black fixation square on a grey background (0.5 x 0.5 cm) for 1.1 s. A single colored circle (stimulus, diameter 1.4 cm, 1 out of 6 randomly chosen colors with equal luminance) was then presented during 0.25 s at one of 360 randomly chosen angular locations at a fixed radius of 4.5 cm from the center. The stimulus was followed by a randomly chosen delay of 0 (16.67% of trials), 1 (66.67% of trials), or 3 s (16.67% of trials) in which only the fixation dot remained visible (except for 0 s trials, where the stimulus remained visible until the participant started to move the cursor). When the fixation dot changed to the stimulus' color (probe), participants were asked to respond by making a mouse click at the remembered location (response). A white circle indicated the stimulus' radial distance, so participants only had to remember the angular position. After the response, the cursor had to be moved back to the fixation dot to start a new trial (ITI). Participants were instructed to maintain fixation during the fixation period, stimulus presentation, and memory delay and were free to move their eyes during response and when returning the cursor to the fixation dot.

Error and serial dependence analysis

Response errors θ_n^e in trial n were measured as the angular distance between response and target. To exclude errors due to guessing or motor imprecision, we only analyzed responses within an angular distance of 1 radian and a radial distance of 2.25 cm from the stimulus. Further, we excluded trials in which the time of response initiation exceeded 3 s, and trials for which the time between the

previous trial's response probe and the current trial's stimulus presentation exceeded 5 s. In total, 2.0% (median, ctrl, i.q.r. = 0.3%), 1.8% (median, enc, i.q.r. = 0.4%) and 2.0% (median, schz, i.q.r. = 0.3%) of trials per participant were rejected.

We then measured serial dependence as the error in the current trial as a function of the circular distance between the previous and the current trial's target location. Fig. 1c,d,e depict 'folded' serial dependence: We multiplied trial-wise errors θ_n^e by the sign of the previous-current distance, θ_n^d : $\theta_n^{e'} = \theta_n^e * \text{sign}(\theta_n^d)$, and then binned data based on absolute values $|\theta_n^d|$. Errors $\theta_n^{e'}$ were then averaged for each $|\theta_n^d|$ in sliding windows with size $\pi/3$ in steps of $\pi/20$. Positive mean folded errors should be interpreted as attraction towards the previous stimulus and negative mean folded errors as repulsion away from the previous location. In all figures including bias curves, s.e.m. are calculated across pooled trials from all subjects for each group and delay. For visualization, all values were transformed from radians to angular degrees.

EEG recordings and preprocessing

We recorded EEG from 43 electrodes attached directly to the scalp. The electrodes were located at Modified Combinatorial Nomenclature sites Fp1, Fpz, Fp2, AF7, AFz, AF8, F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, A1, T7, C5, C3, Cz, C4, C6, T8, A2, TP7, CP3, CPz, CP4, TP8, P7, P3, Pz, P4, P8, PO7, PO3, POz, PO4, PO8, O1, Oz and O2. Sites were referenced to an average of mastoids A1 and A2 and re-referenced offline to an average of all electrodes. We further recorded horizontal EOG from both eyes, vertical EOG from an electrode placed below the left eye and ECG to detect cardiac artifacts. We used a Brainbox® EEG-1166 EEG amplifier with a .017-100 Hz bandpass filter and digitized the signal at 512 Hz using Deltamed Coherence® software (version 5.1).

EEG data was pre-processed using Fieldtrip (version 20171231) in MATLAB R2017b and R2019a. We excluded outlier trials in which variance or kurtosis across samples exceeded 4 standard deviations from mean variance or kurtosis over trials, respectively. To reduce artifacts in the remaining data, we ran an independent component analysis (ICA) on the trial-segmented data and corrected the signal for blinks, eye movements, and ECG signals, as identified by visual inspection of all components. Data were Hilbert-transformed (using the FieldTrip function "ft_freqanalysis.m") to extract frequencies in the alpha-band (8-12 Hz) and total power was calculated as the squared complex magnitude of the signal. Finally, we excluded trials in which lognormal alpha-power at any electrode exceeded the time-resolved trial average of lognormal alpha-power by more than 4 standard deviations, and trials in which the time-averaged variance across electrodes exceeded the mean variance over trials by more than 4 standard deviations (to increase the stability of trial-wise decoding predictions for different randomly chosen training sets). Excluding rejected trials and trials with 0 s delay, we used $n = 900.63 \pm 103.02$ (ctrl, mean \pm std); $n = 870.44 \pm 141.93$ (enc, mean \pm std); and $n = 833 \pm 153.33$ (schz, mean \pm std) trials per participant. To concatenate data from the two sessions for the same subject, we normalized each session's alpha-power for each electrode separately.

EEG decoding

We used a linear decoder to read out the angular position of the stimulus from the distribution of alpha power across the 43 electrodes. We concatenated consecutive trials and trained the decoder on the stimulus label of the previous trial, to then decode location information throughout the previous and current trial. Trial-wise alpha power for each electrode was modeled as a linear combination of a set of regressors representing the stimulus location in the corresponding trial, $U = WM$, where U is a $J \times K$ matrix of alpha power measured at electrode j in trial k , M is the $N \times K$ design matrix of values for regressor n in trial k , and W is the $J \times N$ weight matrix, mapping the weight for regressor n to electrode j . U and M were given by the experiment, while W was fitted using

least squares. The design matrix M is a set of eight regressors M_n representing expected values for feature n in trial k ⁴⁶. The value of regressor M_n in trial k was determined as $\sin(n\pi/8 - s_k\pi/8 + \pi/2)^T$, where $s_k = [0\dots7]$ indicates which one of eight angular location bins (with width $\pi/8$ radians) included the stimulus shown in trial k .

We measured single-trial stimulus representations using leave-one-out cross-validation, ensuring an equal number of trials from each location bin in the training set (U_t and M_t). We estimated the weight matrix \hat{W} and the design matrix M_k for the left-out trial k , as follows:

$$\hat{W} = U_t M_t^T (M_t M_t^T)^{-1}$$

$$\hat{M}_k = (\hat{W}^T \hat{W})^{-1} \hat{W}^T U_k$$

For each trial and time point, we repeated this analysis 10 times with randomly chosen training sets, and averaged M over all repetitions. Finally, we estimated the predicted angle $\hat{\theta}_k$ as the direction of the vector sum of feature vectors with length M_{nk} pointing at angular location bin centers $b_n = n\pi/8$ ($n = 0\dots7$). Trial-wise decoding strength was then defined as $\cos(\hat{\theta}_k - \theta_k)$. To correlate the decoding strength with behavioral biases (Figs. 3c, 4e-g), we increased the stability of trial-wise measures by training the decoder on temporally averaged data over a 200-ms window in mid-delay (0.65 - 0.85 s, orange marks in Figs. 3b,g,h, 4d). We then separated trials as high-decoding (>75th percentile) or low-decoding (<75th percentile) during delay (Fig. 3c) or respectively, during fixation (Fig. 4e-g), and calculated the difference between absolute errors (Fig. 3c) or respectively, between bias curves (Fig. 4e-g) separately for each set of trials. The difference in bias strength between two groups of trials was calculated for previous-current distances between 0° and 90° .

To explore the temporal generalization of the mnemonic and the response code over time, we trained decoders on the previous stimulus label in independent time windows of the previous and current trial, and tested them in all time points of previous delay and response (Fig. 3d-f, 4a-c, from 0.25 s to 1.25 s after previous stimulus onset and from -0.25 s to 0.25 s after the previous response) and current fixation periods (Fig. 4a-c, -1.25 s to 0.25 s after the stimulus onset of the current trial). For decoding matrices, we averaged training and test data over independent windows of 50 samples (~97.77 ms). The high-resolution time course of the mnemonic code (Figs. 3g, 4d) were obtained by training the decoder on averaged data from 0.65 - 0.85 s after previous stimulus onset (dashed lines in decoding matrices), and by testing on averaged data from five samples (~9.77 ms) through consecutive trials. For all timeseries and cross-temporal matrices, significance was assessed with 1-sample or 2-sample cluster permutation tests ⁴⁷ with 1,000 permutations, implemented in the Python “mne” package ⁴⁸.

Neural network simulations

We simulated consecutive pairs of trials in a spiking neural network model of prefrontal cortex implemented in Brian2 ⁴⁹. $N_E = 1024$ excitatory and $N_I = 256$ inhibitory leaky integrate-and-fire neurons were connected all-to-all via synapses governed by NMDAR-, AMPAR-, and GABA_AR-dynamics, as described in ref. ⁵⁰. All connection strengths of all-to-all connections were constant, except for recurrent excitatory connections, which were modulated depending on the distance in preferred location of presynaptic and postsynaptic neurons: $W_{ij}^{EE} = J(\theta_i - \theta_j)$, where J is a Gaussian function (centered at $\mu = 0$ with $\sigma = 14.4$ degrees) plus a constant, tuned so that $\sum_j J(\theta_i - \theta_j) = N_E$ and $J(0) = 1.63$. As a result, neurons with similar preferred locations had 1.63 stronger weights than the average weight. For equations describing network dynamics, please refer to ref. ¹⁶.

Moreover, connections between excitatory neurons were plastic (Fig. 2a): AMPAR and NMDAR synaptic variables w_{ij} characterize the synaptic weight between neuron j and neuron i . Upon synchronized pre- and postsynaptic spiking, w_{ij} was slightly enhanced by an amount Δ_w that depended on the relative spike times of neuron j and i to simulate an increase in probability of glutamate release⁵¹:

$$w_{ij} = w_{ij} + \Delta_w(t_j - t_i) \geq 1.$$

The associative nature of this rule was determined by a potentiation function that required synchronization within a specific temporal window:

$$\Delta_w(t_j - t_i) = P \exp(-|t_j - t_i|/\tau_\Delta),$$

with potentiation factor $P = 0.00022$ (to simulate STP in healthy controls), $P = 0.00012$ (encephalitis), or $P = 0.00002$ (schizophrenia), and $\tau_\Delta = 20$ ms (Fig. 2b). Changes were sustained (did not decay with time), but synapses depotentiated based on presynaptic activity²⁸: at each presynaptic spike:

$$w_{ij} = w_{ij} - 0.04 * (w_{ij} - 1).$$

Dynamics in network connections are described in more detail in ref. ¹⁶.

We simulated 1,000 pairs of consecutive trials with independent randomized stimulus locations. We then calculated rate tuning as the difference between firing rates of neurons selective for the presented location (in trial $n-1$) and neurons selective for the opposite location. Simulations started with a stimulus presentation at 0° (trial $n-1$) for 0.25 s. After the input was removed, a delay of 1 s followed. A negative input to the whole network during 0.25 s simulated the response and removed stimulus-associated neural activity. After an ITI of 2.75 s, a second stimulus (trial n) was delivered at a random location for 0.25 s. 1.1 s before the second stimulus, a transient excitatory drive (0.5 s) was delivered to all excitatory neurons in the network.

References

1. Conway, A. R. A., Kane, M. J. & Engle, R. W. Working memory capacity and its relation to general intelligence. *Trends Cogn Sci (Regul Ed)* **7**, 547–552 (2003).
2. Baddeley, A. Working memory. *Science* **255**, 556–559 (1992).
3. Millan, M. J. *et al.* Cognitive dysfunction in psychiatric disorders: characteristics, causes and the quest for improved therapy. *Nat. Rev. Drug Discov.* **11**, 141–168 (2012).
4. Goldman-Rakic, P. S. Working memory dysfunction in schizophrenia. *J. Neuropsychiatry Clin. Neurosci.* **6**, 348–357 (1994).
5. Dalmau, J. *et al.* Paraneoplastic anti-N-methyl-D-aspartate receptor encephalitis associated with ovarian teratoma. *Ann. Neurol.* **61**, 25–36 (2007).
6. Graus, F. *et al.* A clinical approach to diagnosis of autoimmune encephalitis. *Lancet Neurol.* **15**, 391–404 (2016).
7. Javitt, D. C. Glutamatergic theories of schizophrenia. *Isr. J. Psychiatry Relat. Sci.* **47**, 4–16 (2010).
8. Olney, J. W., Newcomer, J. W. & Farber, N. B. NMDA receptor hypofunction model of schizophrenia. *J. Psychiatr. Res.* **33**, 523–533 (1999).
9. Wang, X. J. Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *J. Neurosci.* **19**, 9587–9603 (1999).
10. Morgan, C. J. A. & Curran, H. V. Acute and chronic effects of ketamine upon human memory: a review. *Psychopharmacology (Berl)* **188**, 408–424 (2006).
11. Gilmour, G. *et al.* NMDA receptors, cognition and schizophrenia--testing the validity of the NMDA receptor hypofunction hypothesis. *Neuropharmacology* **62**, 1401–1412 (2012).
12. Murray, J. D. *et al.* Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. *Cereb. Cortex* **24**, 859–872 (2014).
13. Cano-Colino, M. & Compte, A. A computational model for spatial working memory deficits in schizophrenia. *Pharmacopsychiatry* **45 Suppl 1**, S49-56 (2012).
14. Catts, V. S., Lai, Y. L., Weickert, C. S., Weickert, T. W. & Catts, S. V. A quantitative review of the postmortem evidence for decreased cortical N-methyl-D-aspartate receptor expression levels in schizophrenia: How can we link molecular abnormalities to mismatch negativity deficits? *Biol. Psychol.* **116**, 57–67 (2016).
15. Hughes, E. G. *et al.* Cellular and synaptic mechanisms of anti-NMDA receptor encephalitis. *J. Neurosci.* **30**, 5866–5875 (2010).
16. Stein, H. *et al.* Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia. *Nat. Commun.* **11**, 4250 (2020).
17. Kiyonaga, A., Scimeca, J. M., Bliss, D. P. & Whitney, D. Serial Dependence across Perception, Attention, and Memory. *Trends Cogn Sci (Regul Ed)* **21**, 493–497 (2017).
18. Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* **61**, 331–349 (1989).
19. Kilpatrick, Z. P. Synaptic mechanisms of interference in working memory. *Sci. Rep.* **8**, 7879 (2018).
20. Bliss, D. P. & D'Esposito, M. Synaptic augmentation in a cortical circuit model reproduces serial dependence in visual working memory. *PLoS ONE* **12**, e0188927

- (2017).
21. Akrami, A., Kopec, C. D., Diamond, M. E. & Brody, C. D. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature* **554**, 368–372 (2018).
 22. Barbosa, J. *et al.* Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nat. Neurosci.* **23**, 1016–1024 (2020).
 23. Zucker, R. S. & Regehr, W. G. Short-term synaptic plasticity. *Annu. Rev. Physiol.* **64**, 355–405 (2002).
 24. Markram, H., Wang, Y. & Tsodyks, M. Differential signaling via the same axon of neocortical pyramidal neurons. *Proc Natl Acad Sci USA* **95**, 5323–5328 (1998).
 25. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic theory of working memory. *Science* **319**, 1543–1546 (2008).
 26. Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C. & Gerstner, W. Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* **334**, 1569–1573 (2011).
 27. Castro-Alamancos, M. A. & Connors, B. W. Short-term synaptic enhancement and long-term potentiation in neocortex. *Proc Natl Acad Sci USA* **93**, 1335–1339 (1996).
 28. Volianskis, A. *et al.* Different NMDA receptor subtypes mediate induction of long-term potentiation and two forms of short-term potentiation at CA1 synapses in rat hippocampus in vitro. *J Physiol (Lond)* **591**, 955–972 (2013).
 29. Wimmer, K., Nykamp, D. Q., Constantinidis, C. & Compte, A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neurosci.* **17**, 431–439 (2014).
 30. Foster, J. J., Sutterer, D. W., Serences, J. T., Vogel, E. K. & Awh, E. The topography of alpha-band activity tracks the content of spatial working memory. *J. Neurophysiol.* **115**, 168–177 (2016).
 31. Medendorp, W. P. *et al.* Oscillatory activity in human parietal and occipital cortex shows hemispheric lateralization and memory effects in a delayed double-step saccade task. *Cereb. Cortex* **17**, 2364–2374 (2007).
 32. King, J. R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn Sci (Regul Ed)* **18**, 203–210 (2014).
 33. Stokes, M. G. *et al.* Dynamic coding for cognitive control in prefrontal cortex. *Neuron* **78**, 364–375 (2013).
 34. Starc, M. *et al.* Schizophrenia is associated with a pattern of spatial working memory deficits consistent with cortical disinhibition. *Schizophr. Res.* **181**, 107–116 (2017).
 35. Hahn, B., Robinson, B. M., Leonard, C. J., Luck, S. J. & Gold, J. M. Posterior parietal cortex dysfunction is central to working memory storage and broad cognitive deficits in schizophrenia. *J. Neurosci.* **38**, 8378–8387 (2018).
 36. Driesen, N. R. *et al.* Impairment of working memory maintenance and response in schizophrenia: functional magnetic resonance imaging evidence. *Biol. Psychiatry* **64**, 1026–1034 (2008).
 37. First, M. B., Spitzer, R. L., Gibbon, M. & Williams, J. B. W. *Structured Clinical Interview for DSM-IV Axis I Disorders, Clinician Version (SCID-CV)*. (American Psychiatric Press, Inc., 1996).
 38. Dalmau, J. *et al.* Anti-NMDA-receptor encephalitis: case series and analysis of the

- effects of antibodies. *Lancet Neurol.* **7**, 1091–1098 (2008).
39. Titulaer, M. J. *et al.* Treatment and prognostic factors for long-term outcome in patients with anti-NMDA receptor encephalitis: an observational cohort study. *Lancet Neurol.* **12**, 157–165 (2013).
 40. Finke, C. *et al.* Cognitive deficits following anti-NMDA receptor encephalitis. *J. Neurol. Neurosurg. Psychiatr.* **83**, 195–198 (2012).
 41. Kay, S. R., Fiszbein, A., Vital-Herne, M. & Fuentes, L. S. The Positive and Negative Syndrome Scale--Spanish adaptation. *J. Nerv. Ment. Dis.* **178**, 510–517 (1990).
 42. Colom, F. *et al.* [Spanish version of a scale for the assessment of mania: validity and reliability of the Young Mania Rating Scale]. *Med Clin (Barc)* **119**, 366–371 (2002).
 43. Ramos-Brieva, J. A. & Cordero-Villafafila, A. A new validation of the Hamilton Rating Scale for Depression. *J. Psychiatr. Res.* **22**, 21–28 (1988).
 44. Bobes, J., Portilla, M. P. G., Bascarán, M. T., Saiz, P. A. & Bousoño, M. *Banco de instrumentos básicos para la práctica de la psiquiatría clínica.* (Psiquiatría Editores S.L., 2004).
 45. Woods, S. W. Chlorpromazine equivalent doses for the newer atypical antipsychotics. *J. Clin. Psychiatry* **64**, 663–667 (2003).
 46. Brouwer, G. J. & Heeger, D. J. Decoding and reconstructing color from responses in human visual cortex. *J. Neurosci.* **29**, 13992–14003 (2009).
 47. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007).
 48. Gramfort, A. *et al.* MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* **7**, 267 (2013).
 49. Stimberg, M., Brette, R. & Goodman, D. F. Brian 2, an intuitive and efficient neural simulator. *elife* **8**, (2019).
 50. Compte, A., Brunel, N., Goldman-Rakic, P. S. & Wang, X. J. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb. Cortex* **10**, 910–923 (2000).
 51. Volianskis, A. *et al.* Long-term potentiation and the role of N-methyl-D-aspartate receptors. *Brain Res.* **1621**, 5–16 (2015).
 52. Bliss, D. P., Sun, J. J. & D'Esposito, M. Serial dependence is absent at the time of perception but increases in visual working memory. *Sci. Rep.* **7**, 14739 (2017).

Chapter 4

Summary of Results

In this thesis, I investigated different synaptic and circuit mechanisms of working memory, how their interaction produces specific working memory biases, and how their disruption in psychiatric or neurological disease can contribute to abnormal working memory function.

In Chapter 3.1, I showed that PFC represents working memory contents not only in spiking, persistent activity, but also shows signatures of imprinted, synaptic traces of working memory. These traces can hold contents for an extended period of time, such as an ITI, without the need for firing rate-based maintenance. From locally facilitated synapses, stable working memory representations could be reactivated through unspecific network inputs, a result found in monkey PFC and indirectly in human EEG, and explained by a circuit model of PFC that exhibits bistability (i.e., stable, persistent activity) and is supported by a STP mechanism. Finally, memory reactivations as observed in monkey PFC, human EEG, and elicited with prefrontal TMS in humans increased systematic biases towards previous memories. These findings demonstrate the behavioral relevance as well as the prefrontal locus of the discussed mechanisms.

Chapter 3.2 builds up on the findings from Chapter 3.1. I designed experiments parallel to Chapter 3.1 to test the impact of NMDAR dysfunction on working memory precision and systematic serial biases in patients with anti-NMDAR encephalitis and patients with schizophrenia, and compared their data to that of healthy controls. I found that working memory precision in both patient groups was unaffected, but serial biases were drastically reduced in encephalitis patients, and completely disrupted in patients with schizophrenia. Moreover, I found that biases normalized in patients with encephalitis, a sign of their relation to clinically relevant processes. I then modeled these findings in a prefrontal circuit model similar to the model developed in Chapter 3.1. By disrupting different

NMDAR-related parameters, I showed that perturbations in E/I balance through reduced NMDAR-mediated currents cannot explain findings from patients. In contrast, reduced short-term potentiation successfully disrupted between-trial memory traces and the emergence of serial biases in the model.

In Chapter 3.3, I tested whether neural reactivations of working memory codes, a mechanism of serial dependence identified in Chapter 3.1, would be disrupted in patients' EEG. This question was based on the hypothesis of disrupted STP, as proposed in the circuit model from Chapter 3.2. The findings in Chapter 3.3 confirm our hypothesis: While in healthy controls, memory representations from previous trials were reactivated during the ITI, and subsequently influenced behavior, no such reactivation occurred in patients with anti-NMDAR encephalitis or schizophrenia.

Chapter 5

Discussion

5.1 How do these findings impact our understanding of working memory?

In Chapter 1.1 and 1.3, I have introduced the concepts of persistent delay activity and short-term plasticity (STP), and the ways the two mechanisms are thought to be involved in working memory. Specifically, persistent delay activity has long been accepted as a self-sufficient mechanism that can maintain information in an active state for several seconds (Goldman-Rakic 1995), a hypothesis supported by experimental (Fuster and Alexander 1971; Kubota and Niki 1971; Funahashi et al. 1989; Constantinidis et al. 2001; Wimmer et al. 2014) as well as theoretical work (Compte et al. 2000; Amit and Brunel 1997; Lim and Goldman 2013). In contrast, there has been steadily accumulating in-vitro evidence for different forms of STP in neocortical areas (Hempel et al. 2000; Wang et al. 2006; Castro-Alamancos and Connors 1996). The implication of STP in cognition or working memory, albeit an attractive idea (Miller et al. 2018), has been postulated in theoretical work (Mongillo et al. 2008; Kilpatrick 2018), but has since remained without experimental proof. In Chapter 3.1, we have now demonstrated that a mechanism which temporally and locally enhances neural excitability in PFC, potentially a form of STP, co-exists with persistent neural spiking during the execution of a working memory task. We have shown that this enhancing mechanism is related to the retention of memory contents in persistent activity in an interactive way, and our results show that it plays a functional role in cognition and behavior.

Crucially, Chapter 3.1 does not claim that STP is the main mechanism responsible for delay-period working memory maintenance, like it has been argued by previous studies (e.g. Wolff et al. 2017; Rose et al. 2016; Stokes 2015). In fact, our data does not support the absence of memory codes during short delays, as memories were decodable from a small number of prefrontal neurons throughout the delay. Rather, we adopt a perspective on STP that has its co-existence with persistent delay activity at its core, similar to previous theoretical accounts (Carter and Wang 2007; Itskov et al. 2011; Hansel and Mato 2013; Barbieri and Brunel 2008; Seeholzer et al. 2019; York and van Rossum 2009; Pereira and Wang 2015; Mongillo et al. 2012; Kilpatrick 2018; Yoon et al. 2020). In accordance with the

experimental data shown in Chapter 3.1, our computational model solves the problem of working memory through persistent activity. Delay activity in our task only ceases once working memory contents are recalled and a response is made. After delay-active neurons return to their baseline firing rates, a plastic increase in neuronal excitability becomes evident in the spiking patterns of those neurons (Chapter 3.1, Figure 3)¹⁸, although STP probably already emerges during the delay: Persistent delay firing with rates of ~ 20-60 Hz could represent a short form of tetanic stimulation, similar to the average length and rates of stimulation protocols used in in-vitro studies of short-term potentiation (Erickson et al. 2010; Volianskis et al. 2013; Castro-Alamancos and Connors 1996; Malenka 2002). With regard to the functional role of STP, our study does not take a prominent stance, given that our experiment is not suited to test whether or how working memory performance could benefit from STP. Previous work has postulated that STP mechanisms could replace delay firing (Mongillo et al. 2008; Stokes 2015; Barak and Tsodyks 2007; Fiebig and Lansner 2017), a perspective that we reject for our paradigm based on our finding of a sustained memory code during delay in monkey PFC and human EEG (Chapter 3.1, Figures 1c, 2c), or alternatively stabilize memory representations (Carter and Wang 2007; Itskov et al. 2011; Seeholzer et al. 2019; York and van Rossum 2009; Pereira and Wang 2015; Kilpatrick 2018). It is possible that STP simply emerges as a by-product of delay firing, without serving one specific purpose. In some contexts, this might be beneficial for working memory (i.e., stabilizing delay codes), while in others, it could introduce working memory errors (see Chapter 1.2), such as serial biases.

By characterizing the interplay of STP and persistent activity, we have elucidated the circuit basis of serial dependence in working memory in Chapter 3.1. Our modeling provides a parsimonious framework for how the dynamics in a single, memory-maintaining area can produce at the same time reliable working memory codes, and small, but behaviorally relevant distortions of these codes. In our framework, we do not consider the contribution of inputs from other brain regions to the generation of attractive serial biases: This conceptual choice is based on the finding that serial biases are absent in trials with very short delays, and increase over the course of the delay (see Bliss et al. (2017) and Chapter 3.2), pointing to an origin in memory-maintaining circuits. There might however be complex interactions between several brain regions that could alternatively cause biases in PFC. For example, Akrami et al. (2018) showed that in rat PPC, previous-trial information is

¹⁸ For this finding, the absence of persistent delay spiking for determining STP in data is important, as firing rates might confound cross-correlations measured to detect locally enhanced synapses, as discussed in Chapter 3.1.

represented transiently during ITIs, while the same area did not code for memory items during the working memory delay. Following this observation, it is possible that memories are transferred from PFC to a different, trial-history tracking area after the response, to then influence new memories in the next trial. This possibility is explicitly discussed in Chapters 3.2 and 3.3, as it would influence the interpretation of disrupted serial dependence in patients: Disrupted long-range connectivity could cause a disruption in bi-directional information transfer between memory circuits and trial-history tracking areas.

While the findings of this thesis strongly suggest the existence of a synaptic mechanism that enhances excitability, co-exists and interacts with activity-based memory maintenance, and influences cognition, it is still highly speculative what this mechanism could look like. In this sense, choosing one over the other implementation for the modeling aspects of this thesis should not be understood as a decision with absolute finality. In Chapter 3.1, we used a plasticity rule describing short-term facilitation, similar to refs. (Markram et al. 1998; Mongillo et al. 2008). Short-term facilitation enhances the efficacy of postsynaptic responses non-associatively, as a result of residual calcium in presynaptic terminals subjected to repeated activation (Catterall et al. 2013). Presynaptic calcium levels remain augmented for ~ 1 s after neural activity of the presynaptic cell ceases, determining the relatively fast decay of this plasticity mechanism, compared to longer-lasting short-term potentiation (Zucker and Regehr 2002). The implication of calcium-triggered short-term facilitation in working memory and other computations performed by the PFC is plausible, given that facilitation is widely found in this cortical region and overcomes short-term depression (Hempel et al. 2000; Wang et al. 2006). Potentially, this form of short-term facilitation could also be involved in schizophrenia and explain behavioral findings from Chapter 3.2, given evidence from genetic studies that found calcium channel dysfunction in schizophrenia (Nanou and Catterall 2018).

In contrast, we explicitly modeled an NMDAR-dependent form of STP to explain findings from Chapter 3.2 and 3.3. This choice was based on the assumption that (working) memory alterations in both anti-NMDAR encephalitis and schizophrenia should be related to NMDAR dysfunction, and consequently on the consideration of different synaptic sites and mechanisms that could be affected by NMDAR hypofunction. As set forth in Chapter 1.3, these effects could include processes on different time scales: On the one hand, the NMDAR plays a role in prolonging the postsynaptic response through long-lasting EPSCs, which can contribute to single-cell and network bistability in PFC (Wang 1999; Lisman et al.

1998). On the other hand, the NMDAR regulates different forms of associative potentiation that are distinguished by the time constant of their decay (Volianskis et al. 2013; Erickson et al. 2010). There is evidence for an NMDAR-dependent, fast-decaying component of LTP (with time constant $\tau = 7$ min at a stimulation rate of 0.067 Hz in vitro (Volianskis et al. 2015), and a presumably much faster decay at higher stimulation rates and in vivo environmental conditions, e.g. temperature), including in neocortex (Castro-Alamancos and Connors 1996), which is called early-, transient-LTP or short-term potentiation (described in more detail in Chapter 1.3). Although not nearly as well-established as short-term facilitation, it is possible that an NMDAR-dependent plasticity mechanism is at play in the generation of serial dependence. In turn, it is also possible that forms of STP that do not (directly) depend on NMDARs are affected in patients through some different mechanisms than NMDAR dysfunction (for example through calcium channel dysfunction, as mentioned above). Before further experiments will investigate these questions in pharmacological studies, the conclusions about the specific STP mechanism that maintains activity-independent working memory traces will remain a subject of mere theoretical deliberation.

Finally, this thesis explores the boundaries between memory functions on different timescales, introduced in Chapter 1.1: Possibly, the cognitive processes called working memory and long-term memory are not as neatly separated as classically assumed in the laboratory context. There likely exists a multitude of memory processes between those two extremes, one of which is reflected by the serial dependence that relies on a trial-to-trial information carry-over, implemented in plasticity-based neural codes. In this sense, Chapter 3.1 shows how memory traces with distinct biological substrates can co-exist and influence each other, and Chapters 3.2 and 3.3 show how pathological brain conditions can selectively disrupt some, but not other computations, even as they take place in the same neural circuit. In the context of disease, this finding could impact the way we understand clinical reports of *working memory deficits* in schizophrenia or anti-NMDAR encephalitis, as discussed later in this chapter. For basic research in neuroscience, thinking about the boundaries between working memory, “middle-term memory”, and long-term memory will advance our understanding of the field. To determine the interactions between memory processes that take place concurrently, but at different timescales is a challenge for future work in neuroscience, and a mechanism-centered view will be key for making progress in these questions.

5.2 Does NMDAR dysfunction not affect working memory maintenance, after all?

To establish the clinical relevance and the broader context of the findings reported in Chapter 3.2 and 3.3, it is important to consider their integration with previous literature. While we did not identify precision-related working memory deficits in the patients included in our study, we measured a new behavioral alteration in visual working memory¹⁹ that has not been reported before in any clinical group. Therefore, there are two questions to ask: First, why did we fail to find memory effects in patients with anti-NMDAR encephalitis and schizophrenia that were previously reported in patients with schizophrenia, and pharmacological studies with NMDAR antagonists? Second, how do our new findings fit in with previously reported cognitive alterations in anti-NMDAR encephalitis, schizophrenia, and NMDAR dysfunction more generally, and what do they tell us about the clinical picture of those conditions? In Chapter 5.2, I will discuss the first of these two questions, to then interpret our novel findings in a clinical context in Chapter 5.3.

In our behavioral study in patients, we expected to observe a delay-dependent reduction in spatial working memory precision. This hypothesis was primarily based on previous studies that have reported such deficits in patients with schizophrenia (Badcock et al. 2008; Driesen et al. 2008; Starc et al. 2017; Gold et al. 2020; Park and Holzman 1992) and anti-NMDAR encephalitis (Finke et al. 2012), and supported further by early modeling work that postulates a central role of the NMDAR for persistent delay firing (Wang 1999; Lisman et al. 1998; Compte et al. 2000), over pharmacological manipulations that produced behavioral and physiological results in line with this modeling (Driesen et al. 2013; Wang et al. 2013). In addition, all of our modeling scenarios predict decreased memory precision for the parameters chosen to represent patient groups (Chapter 3.2, Figure 3a-c), although

¹⁹ Note however that for autism spectrum disorder (ASD), (Lieder et al. 2019) found a very similar effect in an auditory delayed comparison task: Auditory memories in subjects with ASD were less biased towards previously remembered frequencies, compared to healthy controls. Moreover, this effect showed a similar “distance-dependent” profile of attraction as serial dependence in visual working memory tasks.

direct manipulations of NMDAR-mediated conductance affected precision to a stronger extent than manipulations of short-term potentiation. Moreover, clinical reports of deficits in neuropsychological working memory tasks in both patient groups strengthened this hypothesis (Finke et al. 2012; Barch and Ceaser 2012), although neuropsychological tasks used to assess working memory often differ from the task used in this thesis in terms of their complexity, typically by requiring the manipulation of memory items (e.g., reverse or ordered recall of encoded items), but also in terms of the delay length.

So why did both groups of patients perform the memory task without measurable increased difficulty or imprecision compared to healthy controls, as reported in Chapter 3.2? Our first intuition was to correct reports for serial biases - given that this source of imprecision (see Chapter 1.2) was absent in patients, and especially in long memory delays, the imprecision of their reports could have been underestimated, when compared to controls. However, correcting for biases did not affect our result. Also, the extremely simplistic task design did not host any other factors that might be masking memory deficits in patients: Any difference in task strategy or general task processing would have affected overall accuracy, but not in a delay-dependent fashion (e.g., patients might have been more motivated than healthy controls). In turn, it is possible that our extremely simple task was indeed too simple to measure memory impairment. Indeed, when looking closer at the literature on spatial working memory deficits in schizophrenia, but also pharmacological studies with ketamine in humans, the picture is less clear than initially expected (Lee and Park 2005; Gold et al. 2010; Morgan and Curran 2006). As discussed in Chapter 1.3, delay-dependent working memory impairment is not unequivocally found in these conditions, and studies that did identify such a deficit (Starc et al. 2017; Gold et al. 2020) used delays of up to 20 s, which is substantially longer than the delays commonly used to study working memory in laboratory tasks. Other tasks in which working memory deficits are measured include increased task complexity, e.g. in multi-item tasks, distractor tasks, or tasks used in neuropsychological assessments (Lee and Park 2005).

That maintenance deficits are absent from patients' responses due to ceiling effects in task performance is congruent with the reduced EEG memory traces reported in Chapter 3.3, where patients again completed the task with similar performance, but weaker and less stable memory codes. In our task, weak traces did not measurably result in imprecision or forgetting, but again, they might do so in future studies with longer delays or disruptions through distractors or simultaneously held items. Moreover, future work should clarify the

nature and behavioral relevance of less stable, or more dynamic, memory codes in patients with anti-NMDAR encephalitis or schizophrenia. From the analyses presented in Chapter 3.3, it is not clear what the meaning of a dynamic code in EEG is: In the single-neuron literature, the concept of “dynamic codes” is interpreted at the population level, as a sequence of distinct, memory-encoding neural firing patterns within the same cortical area (Murray et al. 2017; Parthasarathy et al. 2019; Meyers et al. 2008), and even as a rapid change in prefrontal tuning properties (Stokes et al. 2013; Spaak et al. 2017). The benefits of such dynamic coding schemes are a matter of ongoing discussion, and might include the possibility to keep track of the timing of task events (Meyers et al. 2008), e.g. in tasks with fixed delay durations or increased temporal complexity (Orhan and Ma 2019).

In contrast, dynamic coding schemes in EEG should be interpreted as shifts between different brain areas that dominate the working memory code (King and Dehaene 2014). In this sense, separating the spatial sources of stable and dynamic working memory representations will be an important step in understanding the qualitatively different delay codes observed in patients and healthy controls. Different approaches could be useful to achieve this: On the one hand, one could separate signals from frontal and parieto-occipital electrodes, and explore the different coding schemes separately in each subset of electrodes, as proposed by Oh et al. (2019). Here, the authors found neurotypical subjects’ working memory codes at frontal electrodes to be more stable in time, and parieto-occipital codes to be rather dynamic and driven by task demands. However, as discussed in Chapter 1.1, this approach is limited by the overlapping contributions of signals from different brain areas that are integrated at an electrode; i.e., signals at frontal electrodes cannot be unambiguously interpreted as coming from frontal cortex. Alternatively, the sources of memory codes at different periods of the trial could be reconstructed via inverse models (Michel and Brunet 2019). The latter option would lead to a spatially more precise and unequivocal solution, and would offer the possibility to interpret dynamic and weaker memory codes in patients. For example, memory codes might be distributed across different cortices, with stable frontal, and less stable parieto-occipital codes that represent a series of processing stages (Mejias and Wang 2019). In patients, stable frontal codes could be disrupted, while dynamic posterior codes might still be able to hold memory in a sufficiently reliable code to perform our simple working memory task.

Finally, it is a challenging endeavour to reconcile the finding of intact, or not measurably affected, working memory performance in patients with the presumed theoretical

importance of the NMDAR for memory-maintaining circuits. In our current understanding of the biophysics of working memory, decreases in NMDAR function both at pyramidal cells or alternatively at inhibitory interneurons should quickly destabilize network bistability and therefore memory maintenance, as demonstrated in Chapter 3.2. One way to explain intact memory performance is to assume that NMDAR dysfunction symmetrically affects both inhibitory and excitatory connections, thereby keeping the overall E/I ratio balanced. However, this solution would shift the network to a more AMPA-dependent regime and affect the conditions under which persistent delay activity is achieved (Wang 1999). Another way to compensate for the dysfunction of the NMDAR would be an increased reliance on other excitatory channels with longer time constants, such as the kainate receptor (Castillo et al. 1997), or an increased reliance on STP mechanisms (a possibility that is incompatible with the results in Chapter 3.2). Alternatively, it is possible that patients tested in our studies have passed the most acute stages of their disease, and greatly recovered their preclinical cortical NMDAR density. In this scenario, NMDAR-dependent memory maintenance in persistent activity would be greatly affected in early and/or more acute stages of the disease as predicted by perturbations of NMDAR-mediated synaptic conductance parameters (Chapter 3.2), but recovered or compensated for in residual stages of the disease. STP deficits in this case could then result as long-lasting residual, possibly compensatory effects for the disturbed synaptic transmission in more acute stages. Note that in fact, the postulated deficits in NMDAR-dependent short-term potentiation affect network bistability and memory maintenance in our model in a more subtle way than E-I imbalance, congruent with our findings. Still, until (electro-)physiological experiments can clarify the existence and synaptic origin of postulated STP deficits during stabilized phases of anti-NMDAR encephalitis and schizophrenia, it remains a matter of speculation how NMDAR dysfunction would selectively affect STP, but not basic synaptic transmission.

5.3 Which aspects of altered cognition does a disruption in serial dependence in anti-NMDAR encephalitis and schizophrenia reflect?

In this final section, I will discuss how our findings of reduced serial dependence and the hypothesized disruption in short-term potentiation could translate into a clinical context. This discussion will include the potential role of serial dependence for cognition in neurotypical subjects, and how the absence of these biases could reflect abnormal cognitive processing in brain disease. I will then discuss the potential of using serial biases as a biomarker for conditions of NMDAR dysfunction, and specifically for the two disorders studied in this thesis. Finally, I will propose future directions for work that could clarify questions which this thesis opens.

The absence of serial dependence from patients' working memory representations is an exciting finding: It shows that in the tested individuals, a basic mechanism of (voluntary or involuntary) information maintenance in working memory circuits is disturbed. At the same time, however, it is difficult to interpret how this alteration translates to clinical symptoms that characterize the two studied diseases. This difficulty is due to our incomplete understanding of the functional role of serial dependence: In neurotypical subjects, serial biases occur in tasks with entirely uncorrelated stimulus sequences (Kiyonaga et al. 2017), where they are detrimental to task performance: As explained in more detail in Chapter 1.2, biases increase the variability of responses and therefore decrease working memory precision. In this sense, biases are non-optimal, and should be avoided if possible, so that the absence of biases in patients' responses is not primarily a deficit, but a *more optimized* way of solving the task. Still, neurotypical subjects do not seem to strive for optimality when it comes to serial biases: In fact, the strength of serial dependence increases over the course of experimental sessions, instead of decreasing (Barbosa and Compte 2020).

To resolve these apparent contradictions, it is useful to think of serial dependence outside of the context of the laboratory tasks it is studied in: It has been proposed that

biases could be useful in real-life situations, where temporally adjacent scenes are normally correlated (Cicchini et al. 2018). In support of this argument, serial dependence is observed across a wide range of stimuli with different levels of complexity and abstraction, such as orientation (Fischer and Whitney 2014), color (Barbosa and Compte 2020), auditory frequencies (Lieder et al. 2019), numerosity (Cicchini et al. 2014), and faces (Lieberman et al. 2014), to only name some. Thus, serial biases could reflect a general processing mode that is beneficial in most contexts, but not in others (Kiyonaga et al. 2017), such as the task used in this thesis. In this interpretation, trial-to-trial serial dependencies could lie at the end of a spectrum of potentially adaptive processing modes that integrate information across different time spans, together with biases that depend on alternating block-wise contexts (Hermoso-Mendizabal et al. 2020), contraction biases to the expected value of an experiment's stimulus distribution (Akrami et al. 2018), slowly fluctuating heuristic (Mochol et al. 2020) or persistent idiosyncratic choice biases (Lebovich et al. 2019).

If one assumes that biases result from a general processing mode that is adaptive in contexts other than the experiment itself, reduced serial dependence could reflect patients' reduced capacity of integrating old and incoming information in working memory. This interpretation is attractive, as it connects our findings to long-standing theories of schizophrenia which place dysfunctional context processing at the root of cognitive and perceptual symptoms in this disease (Hemsley 2005), as discussed in more detail in Chapter 3.2. Our modeling contributes to this reading, by showing that reduced short-term potentiation can perturb information maintenance on timescales necessary for context processing. From the work presented in this thesis, it would be expected that such a dysfunction would impair the performance of patients with schizophrenia and anti-NMDAR encephalitis in tasks that require the integration of information across trials. To test the temporal bounds of the neural mechanism affected in these disorders, stimulus correlations on distinct timescales could be introduced in variations of the task. Such experiments would also rule out an alternative explanation of our findings: Serial biases in neurotypical subjects could be produced volitionally, based on the (misguided) assumption that correlations exist between the stimuli used in our task. If this were the case, decreased task engagement in patients could lead to decreased biases. However, it is not clear why in this scenario of volitional biases, attraction to previous stimuli would only arise in trials with long memory delays.

How useful could serial dependence be for diagnostic purposes, as a so-called “biomarker” for the diagnosis or evolution of anti-NMDAR encephalitis, schizophrenia or other conditions related to NMDAR dysfunction? Again, extensive future work would be needed to answer this question. First of all, it is not well-studied how stable serial dependence is in individual neurotypical subjects or patients. In Chapter 3.2, we showed that on average within a group, the amplitude of bias curves stays unaltered for healthy controls, and increases for recovering anti-NMDAR encephalitis patients. This result is promising for the clinical relevance of the measure, but would need to be confirmed and studied in more detail to establish its validity in individual subjects, before being deemed useful for diagnostic purposes. Similarly, it is widely known that patients with schizophrenia undergo a series of neurobiological processes between preclinical stages of the disease, its first onset, and states that follow homeostasis and neurobiological compensation (Sohal and Rubenstein 2019), and a great biological and symptomatic heterogeneity between different subtypes of schizophrenia, psychotic and residual episodes, and treated and untreated stages (Liang and Greenwood 2015). Yet, in our experiment we tested a subset of patients that does not cover this wide spectrum of conditions systematically. Interindividual heterogeneity even affects biases measured within the three groups assessed in Chapter 3.2 and 3.3, with sometimes repulsive biases in healthy controls, and both strong repulsive and strong attractive biases in encephalitis patients that average out at the group level.

The last and probably most important limitation to using serial dependence as a diagnostic tool lies in the unknown specificity of our findings: We identified a reduction in biases in two clinical groups, but we did not include a “clinical control” for a double dissociation between NMDAR-associated conditions, and conditions not related to NMDAR hypofunction. The only other clinical groups for which serial dependence has been studied are autism and dyslexia, two disorders related to difficulties in context processing (Lieder et al. 2019). The authors found a reduction in serial dependence in autism, but not in patients with dyslexia²⁰, pointing to the specificity of the disrupted mechanism, rather than to a generalized deficit. Still, more studies in clinical populations are needed and the usefulness of serial dependence for the diagnosis of anti-NMDAR encephalitis or schizophrenia is not clear at this point. In turn, if serial dependence was validated as a stable and reliable measure in individual subjects that correlated with longitudinal health markers, it could potentially serve to monitor clinical progress in patients with anti-NMDAR encephalitis.

²⁰ At the same time, patients with dyslexia had reduced contraction biases, which were not present in patients with autism.

Relatedly, we have based our study on the assumption that working memory alterations in anti-NMDAR encephalitis and schizophrenia should be caused by a common, NMDAR-dependent neurobiological mechanism. Yet, biological mechanisms underlying anti-NMDAR encephalitis and schizophrenia greatly differ (Masdeu et al. 2016; Kayser and Dalmau 2014; Oviedo-Salcedo et al. 2018). Therefore, it is important to understand that our hypothesis and the modeling presented in Chapter 3.2 and 3.3 does not imply that we treat anti-NMDAR encephalitis and schizophrenia as one and the same disease. Rather, based on the convergence in symptoms (notably psychosis and neuropsychological symptoms affecting executive functions and memory; Finke et al. 2012) and the frequent initial misdiagnosis (Steiner et al. 2013; Maneta and Garcia 2014), we expect there to be a common substrate on a level that affects cognitive and/or perceptual processing. This common substrate is likely to involve the NMDA receptor, based on the long-standing hypothesis of NMDAR hypofunction in schizophrenia (Olney et al. 1999). Thus, the comparison of these two diseases is not an arbitrary choice for this thesis, but a sustained theme in the ever-expanding literature on anti-NMDAR encephalitis and schizophrenia (Kayser and Dalmau 2014; Masdeu et al. 2016; Weickert and Weickert 2016; Lennox et al. 2012; Maneta and Garcia 2014). It is still conceivable that distinct biological mechanisms lead to the same phenomenon of reduced serial dependence in the two groups, but without further evidence, a common alteration in NMDAR function and short-term potentiation is the most parsimonious interpretation of our findings.

To conclude, it is possible that serial dependence reflects memory processes on timescales that are not typically probed in laboratory tasks of working memory or long term memory, such as the delayed-response task. In patients with anti-NMDAR encephalitis and schizophrenia, this timescale might be relevant for everyday-life problems, and might even be a better reflection of the cognitive problems that determine the specific clinical experience of these patients. Seeing serial dependence as a phenomenon emerging from memory on intermediate timescales, our findings might even reconcile the inconsistent literature on working memory deficits in clinical and neuropsychological assessments of anti-NMDAR encephalitis and schizophrenia, but not in simple working memory tasks with several seconds of undistracted delays. To come back to our initial definition in Chapter 1.1, memory describes the ways by which the brain achieves continuity of information through time. In addition to the maintenance aspect of this definition, this thesis lays out a way by which a continuous stream of information can be integrated in memory, namely

through the merging of single working memory contents, and the neural substrates of short and intermediate timescales on which working memory contents are represented.

Chapter 6

Conclusions

1. Working memory involves more than one mechanism that can maintain item-specific information. In addition to persistent activity of prefrontal cortex neurons, I have demonstrated that information is represented in traces of enhanced excitability between delay-active neurons. Enhanced excitability traces decay more slowly than persistent delay activity and can span several trials.
2. Traces of enhanced excitability, which could result from short-term plasticity mechanisms, and persistent delay firing are interdependent substrates of working memory: During working memory delays, persistent delay firing triggers (plasticity) processes that temporally enhance the connection between delay-active neurons. In turn, memory-specific enhanced connections allow the network to reactivate previous memory codes in persistent neuronal firing.
3. Reactivations of previous memory codes can be observed in monkey prefrontal cortex neurons and human EEG codes, and play a causal role in biasing upcoming memories.
4. Anti-NMDAR encephalitis and schizophrenia, two diseases linked to the dysfunction of the NMDAR, are characterized by a strong reduction in serial biases compared to healthy controls, while working memory precision remains intact. This reduction in serial dependence evidences the disruption of a working memory mechanism on an intermediate time scale.

5. A spiking neural network model that combines persistent delay firing and NMDAR-dependent short-term potentiation reproduces the delay-dependent profile of serial dependence in healthy controls. A decrease in short-term potentiation can qualitatively and quantitatively capture the behavioral effects observed in anti-NMDAR encephalitis and schizophrenia. In contrast, disrupting synaptic conductances through NMDARs and therefore disturbing the balance between cortical excitation and inhibition cannot explain the findings.

6. Reducing NMDAR-dependent synaptic potentiation in a circuit model disrupts the ability of the network to reactivate previous memories in persistent activity. This prediction is confirmed when analyzing the strength of memory code reactivations in human EEG in patients with anti-NMDAR encephalitis and schizophrenia. This finding provides further evidence for the dysfunction of a plasticity-based memory mechanism at an intermediate timescale in anti-NMDAR encephalitis and schizophrenia.

Appendix

A1	Supplementary Material for Chapter 3.1
A2	Supplementary Material for Chapter 3.2
A3	Supplementary Material for Chapter 3.3
A4	Ten Simple Rules for Modern Psychophysics

A1 Supplementary material for Chapter 3.1

This section²¹ contains Supplementary material which has been published alongside the main article presented in Chapter 3.1.

²¹ This section has been published as:

Barbosa, J.*, Stein, H.*, Martinez, R.L., Galan-Gadea, A., Li, S., Dalmau, J., Adam, K.C.S., Valls-Solé, J., Constantinidis, C., & Compte, A. (2020). Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nat Neurosci* 23, 1016–1024. <https://doi.org/10.1038/s41593-020-0644-4>. (*equal contribution)

In the format provided by the authors and unedited.

Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory

Joao Barbosa ^{1,8}, Heike Stein ^{1,8}, Rebecca L. Martinez¹, Adrià Galan-Gadea¹, Sihai Li², Josep Dalmau ^{1,3,4,5,6}, Kirsten C. S. Adam ⁷, Josep Valls-Solé¹, Christos Constantinidis² and Albert Compte ¹✉

¹Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. ²Department of Neurobiology and Anatomy, Wake Forest School of Medicine, Winston-Salem, NC, USA. ³Service of Neurology, Hospital Clínic, Barcelona, Spain. ⁴University of Barcelona, Barcelona, Spain. ⁵ICREA, Barcelona, Spain. ⁶Department of Neurology, University of Pennsylvania, Philadelphia, PA, USA. ⁷Department of Psychology and Institute for Neural Computation, University of California San Diego, La Jolla, CA, USA. ⁸These authors contributed equally: Joao Barbosa and Heike Stein. ✉e-mail: acompte@clinic.cat

Interplay between persistent activity and activity-silent dynamics in prefrontal cortex underlies serial biases in working memory

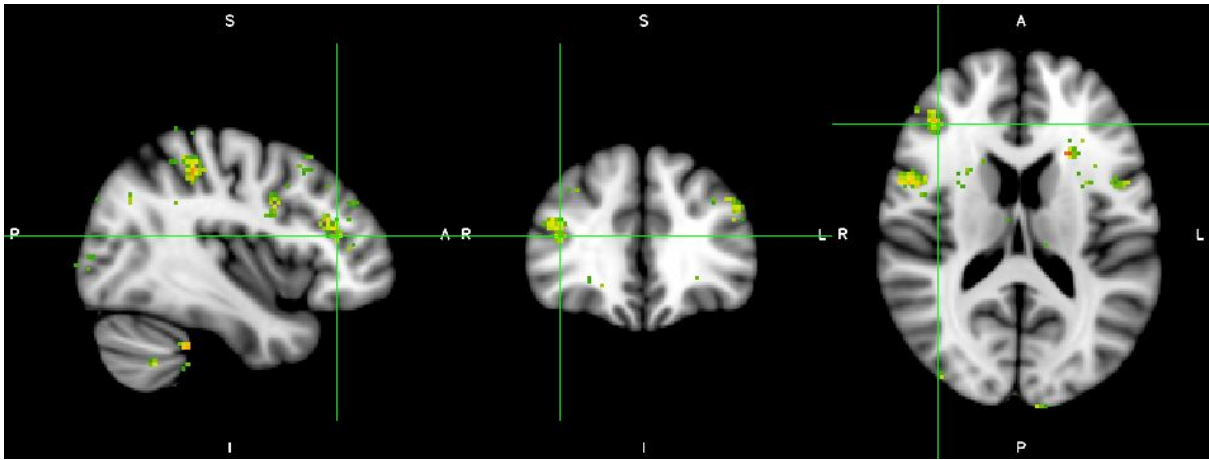
Barbosa J^{1,†}, Stein H^{1,†}, Martinez RL¹, Galan-Gadea A¹, Li S², Dalmau J^{1,3-6}, Adam KCS⁷, Valls-Solé J¹, Constantinidis C², Compte A^{1,*}

1. IDIBAPS, Barcelona, Spain
2. Department of Neurobiology and Anatomy, Wake Forest School of Medicine, Winston-Salem NC, USA
3. Department of Neurology, Hospital Clínic, Barcelona, Spain
4. University of Barcelona, Spain
5. ICREA, Barcelona, Spain
6. Department of Neurology, University of Pennsylvania, USA
7. Department of Psychology and Institute for Neural Computation, University of California San Diego, La Jolla CA, USA

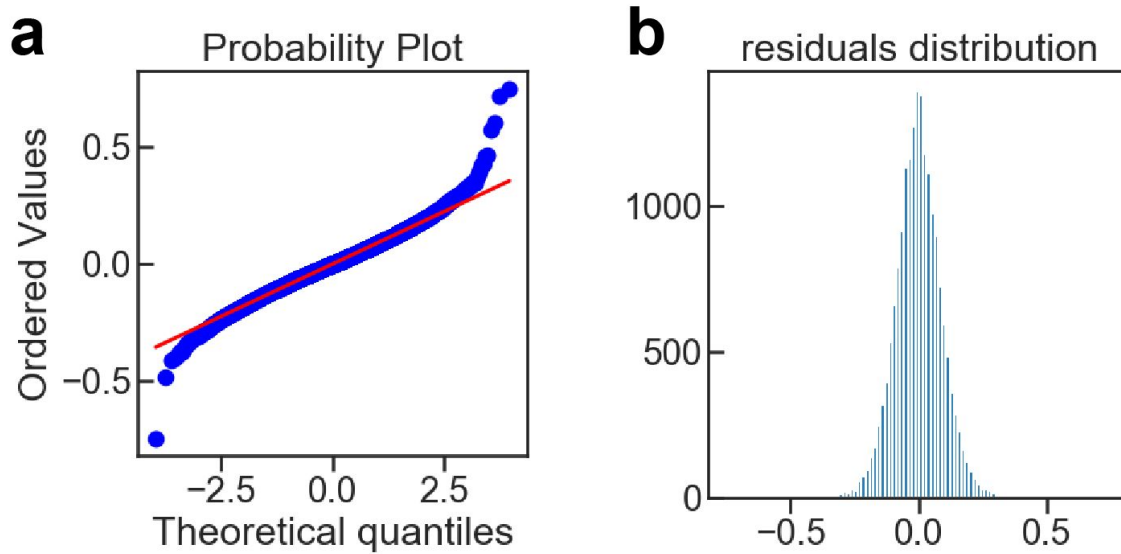
* corresponding author: acompte@clinic.cat

† equal contribution

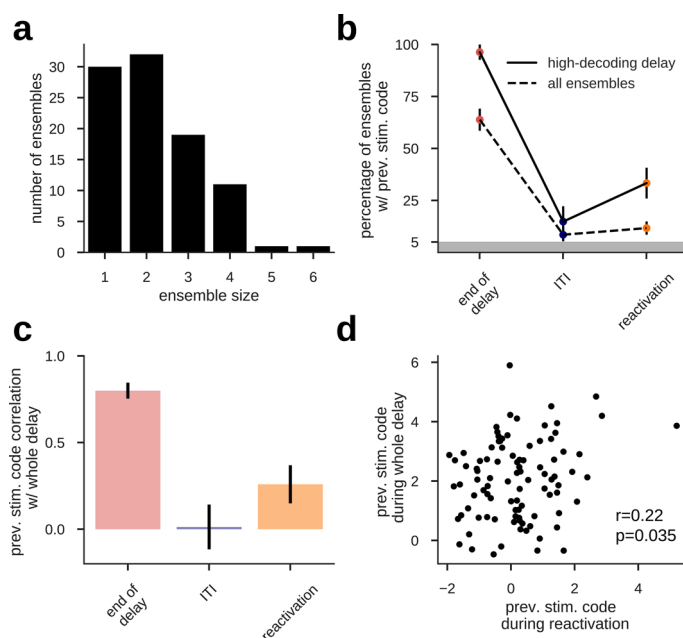
Supplementary Materials: Supplementary Figure 1, Supplementary Figure 2



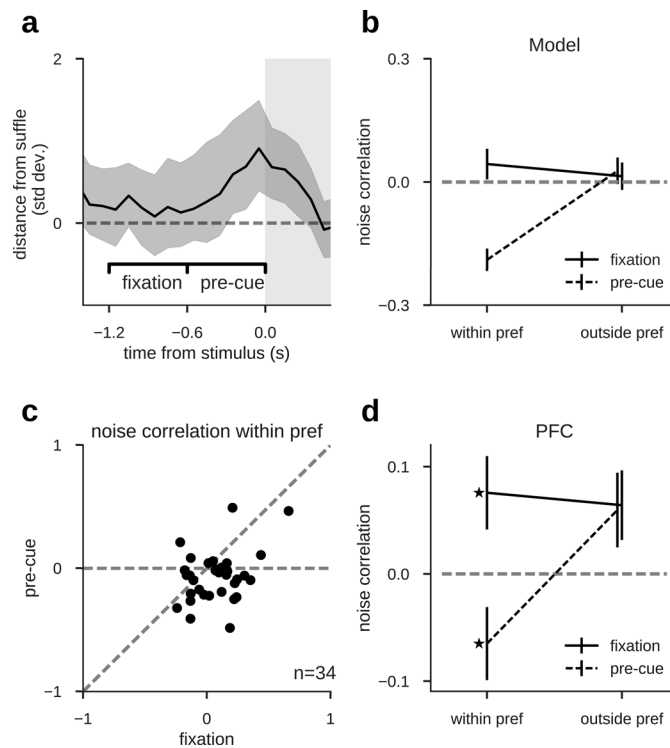
Supplementary Figure 1. MNI coordinates used for TMS stimulation in the right prefrontal cortex. We targeted the center of a functionally defined region of interest (ROI) in right PFC. The ROI was determined as a cluster of activation around MNI coordinates $x = 40$, $y = 34$, $z = 16$ (voxel with maximal activation in right PFC) that was consistently found to be activated in imaging studies investigating “spatial working memory” (custom keyword search in Neurosynth <https://neurosynth.org/> as explained in *Methods*, activation map available as *Supplementary Data*). Here, the ROI is rendered on a 1 mm MNI152 standard in FSL. Before each TMS session, the mask was transformed (12 degrees of freedom) and rendered on each subject’s T1. The depicted ROI was then identified in the subject’s space for TMS stimulation.



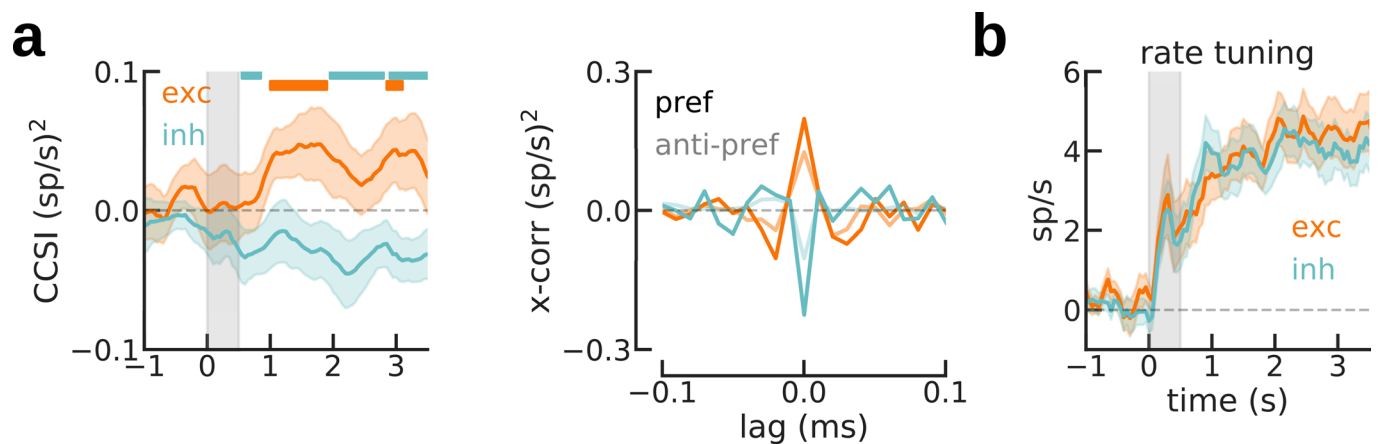
Supplementary Figure 2. qqplot (a) and distribution (b) of residuals for the linear mixed model applied in the TMS data analysis (Fig. 6).



Extended Data Fig. 1 | Consistent decoding accuracy in delay and reactivation links these two representations at the neural ensemble level. a, The size of $n=94$ independent ensembles of simultaneously recorded neurons varies between 1-6. **b**, Fraction of neural ensembles with significant previous stimulus decoding accuracy ($z > 1.96$, see Methods) computed for all ensembles (dashed line) and only for those ensembles with strongest previous stimulus code averaged across the whole delay (see Methods). The incidence of stimulus decoding was significant in delay and reactivation, but not at ITI (two-sided binomial test at $p=0.05$, with $n=94$ and $n=27$ ensembles, for 'all ensembles' and 'highest delay code', respectively). Error bars are bootstrapped \pm s.e.m. **c**, across-ensemble Pearson correlation between delay decoding accuracy (averaged in the entire delay) and decoding accuracy at different time points (two-sided p -values: $6.5e-30$, 0.87 , 0.035 , $n=94$ ensembles). The ensembles with strongest delay code also had stronger decoding during reactivation, demonstrating the neural association between delay representations and reactivations despite absent code in the ITI. Error bars denote \pm s.e.m. computed with a bootstrap procedure. **d**, Individual ensemble values from **c**, orange (Pearson correlation, two-sided $p=0.035$, $n=94$ ensembles).

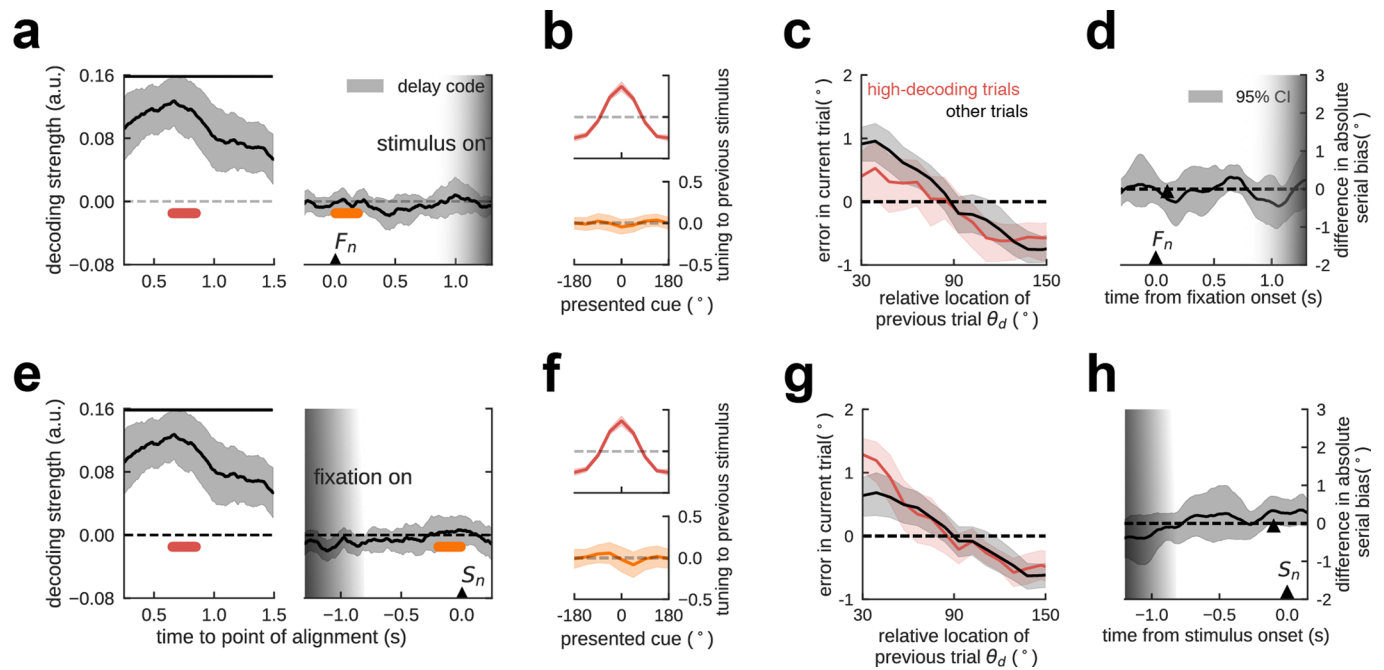


Extended Data Fig. 2 | Noise correlation between pairs of neurons is negative at reactivation, as predicted by the attractor model. Bump-attractor dynamics are characterized by negative pairwise noise correlations for cues presented between the preferred locations (*within pref*) of the two neurons, but not for other cues (*outside pref*)⁶. **a**, Periods used in noise correlation analyses: early (*activity-silent*), and late fixation (*reactivation*; $n=94$ ensembles, zoom-in of Fig. 1c). Error shading, bootstrapped 95% C.I. **b**, In the computational model ($n=1,000$ independent simulations), bump reactivations from subthreshold traces are characterized by negative noise correlations only during reactivation for *within-pref* trials, following the nonspecific input drive (Fig. 4). **c**, Noise correlations of PFC pairs with dissimilar preferred angles ($60^\circ < \Delta\theta < 120^\circ$, $n=34$ pairs) were lower in late than in early fixation for *within-pref* trials (bootstrap test, $p=0.0001$, $n=34$, Cohen's $d=0.61$). **d**, On average, lower noise correlations occurred only during reactivation and in *within-pref* trials (ANOVA *trial condition* \times *time point*, $F(4)=2.5$, $p=0.06$, $n=34$). For *within-pref* trials, noise correlations differed between early and late fixation (bootstrap test, $p=0.0001$, Cohen's $d=0.61$, $n=34$), being negative in late (bootstrap test, $p=0.035$, Cohen's $d=-0.32$, $n=34$), but positive in early fixation (bootstrap test, $p=0.018$, Cohen's $d=0.37$, $n=34$). Correlations were positive in *outside-pref* trials both during late and early fixation (bootstrap test, $p=0.024$ and $p=0.06$, respectively), with no significant difference (two-sided bootstrap test, $p=0.93$, $n=34$). In addition, negative noise correlations diminished when using the previous saccade location rather than the previous stimulus as reference (paired bootstrap test, $p=0.005$, Cohen's $d=-0.47$, $n=34$), suggesting that the bump diffused only during the delay period, but not after the saccade⁶. Unless stated otherwise, all bootstrap tests were one-tailed in the direction of the model predictions in b. All error bars indicate \pm s.e.m.

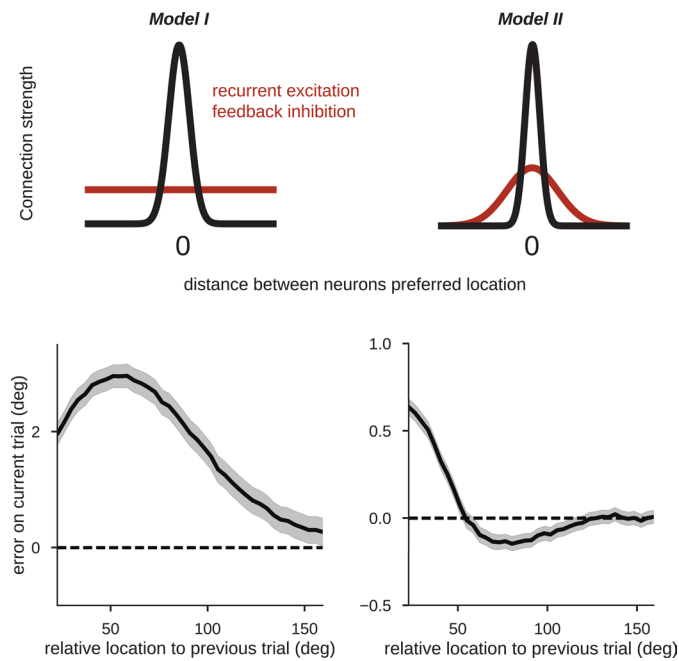


Extended Data Fig. 3 | Stimulus selectivity in both cross-correlation peaks and firing rates during the delay period prevents the isolation of activity-based and activity-silent processes. Same analysis as in Fig. 3, but performed during the current delay period (instead of ITI, Fig. 3) and selecting pref and anti-pref trials based on current stimulus (instead of previous, Fig. 3). Note that these are different trials (no need to be consecutive), so exc ($n=33$ pairs) and inh ($n=21$ pairs) might differ from Fig. 3. **a**, Left, cross-correlation peak selectivity emerged and was sustained in the delay period (left, CCSI as in Fig. 3, computed in centered 500-ms windows sliding in steps of 50 ms) and consisted in enhanced central peaks (troughs) for exc (inh) following a preferred stimulus. Color bars mark the periods where the average CCSI is different from 0 (bootstrapped 95% C.I.) Right, cross-correlation averaged over 0.5–3.5 s. Zero-lag correlation for pref and anti-pref are different in exc ($p=0.03$, $n=33$, two-sided paired bootstrap test) and inh ($p=0.01$, $n=21$, two-sided bootstrap test) conditions. **b**, Firing rate selectivity (pref - anti-pref) also emerges robustly in the delay period for neurons in exc and inh pairs. The selectivity in cross-correlation peaks (CCSI) can therefore be confounded with firing rate selectivity⁷¹ when analyzing data in the delay period. This prevents the unambiguous identification of activity-silent mechanisms in this task period. Our approach of analyzing data in the inter-trial interval, when there is no firing rate selectivity (Fig. 3f), gets around this problem. Gray shading marks the stimulus presentation. In all panels, error-bar shadings indicate \pm s.e.m.

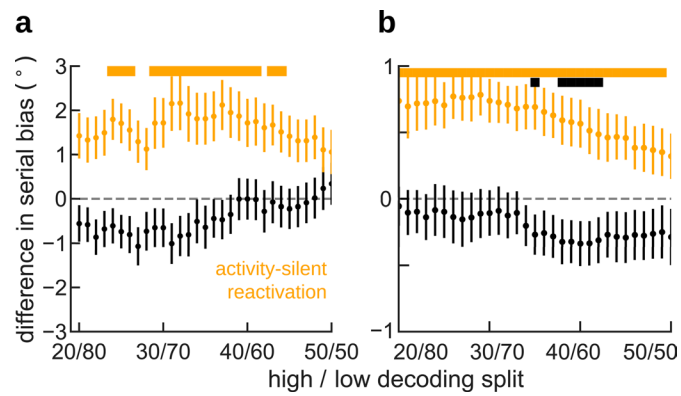
71. de la Rocha, J., Doiron, B., Shea-Brown, E., Josić, K. & Reyes, A. Correlation between neural spike trains increases with firing rate. *Nature* **448**, 802–806 (2007).



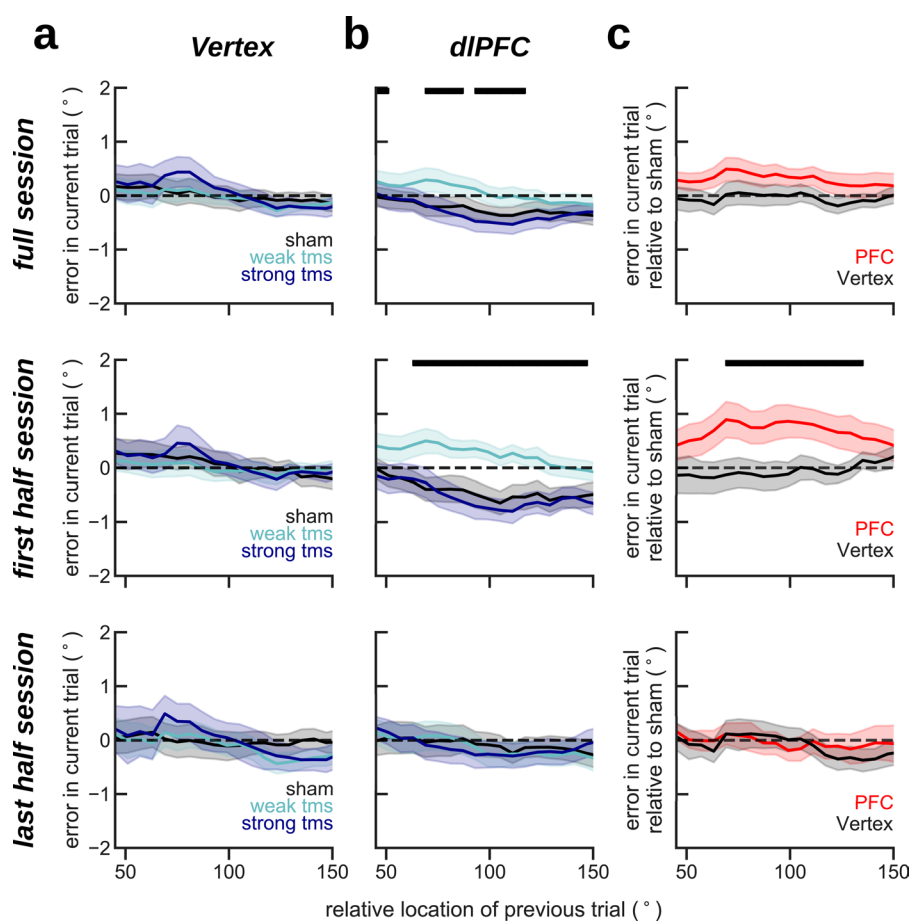
Extended Data Fig. 4 | In a dataset with unpredictable stimulus-onset time, previous item representations were not reactivated in the pre-stimulus period. We conducted the same analysis as in human EEG (Fig. 2) in a previously published dataset ($n=15$ independent subjects for all panels; for experimental details, please refer to the original publication, ref.³³) with unpredictable fixation period durations (range 0.7 s-1.3 s). Decoding analyses were applied separately for data aligned to the onset of fixation (F_n , graded shading indicates range of possible stimulus onset times, upper panels) and aligned to the onset of the stimulus (S_n , graded shading indicates possible fixation onset times, lower panels). **a**, Tuning to previous-trial location (decoder trained in delay, 0.5s - 1.0s after stimulus onset) during previous-trial delay (left, stimulus aligned) vanishes in current-trial fixation (right, fixation onset aligned). No reactivation occurs. **b**, Average tuning reconstruction at different epochs for the delay decoder, indicated in **a**. **c**, Serial dependence separating trials with high (red curve, top quartile) from all other trials' (black curve) decoding accuracy in early fixation (orange in **a**). Unlike in an experiment with predictable stimulus onset (Fig. 5), serial bias did not differ as a function of decoding strength. **d**, Difference in serial biases (Methods) between *high-decoding* and *other* trials were not significant at any time point in fixation. The black triangle marks the center of 0.2 s decoding window for the split in **c**. **e-h**, Parallel results were obtained when the analyses of panels **a-d** were run on data aligned to the time of stimulus onset instead of fixation onset. In **d** and **h**, time courses were smoothed using a squared filter of 5 samples. Periods with significant decoding in **a,e** are marked with black horizontal bars, indicating $p < .001$ in a two-sided bootstrap test. Shading indicates 95% C.I. in **a,d,e,h**, and \pm s.e.m. in **b,c,f,g**.



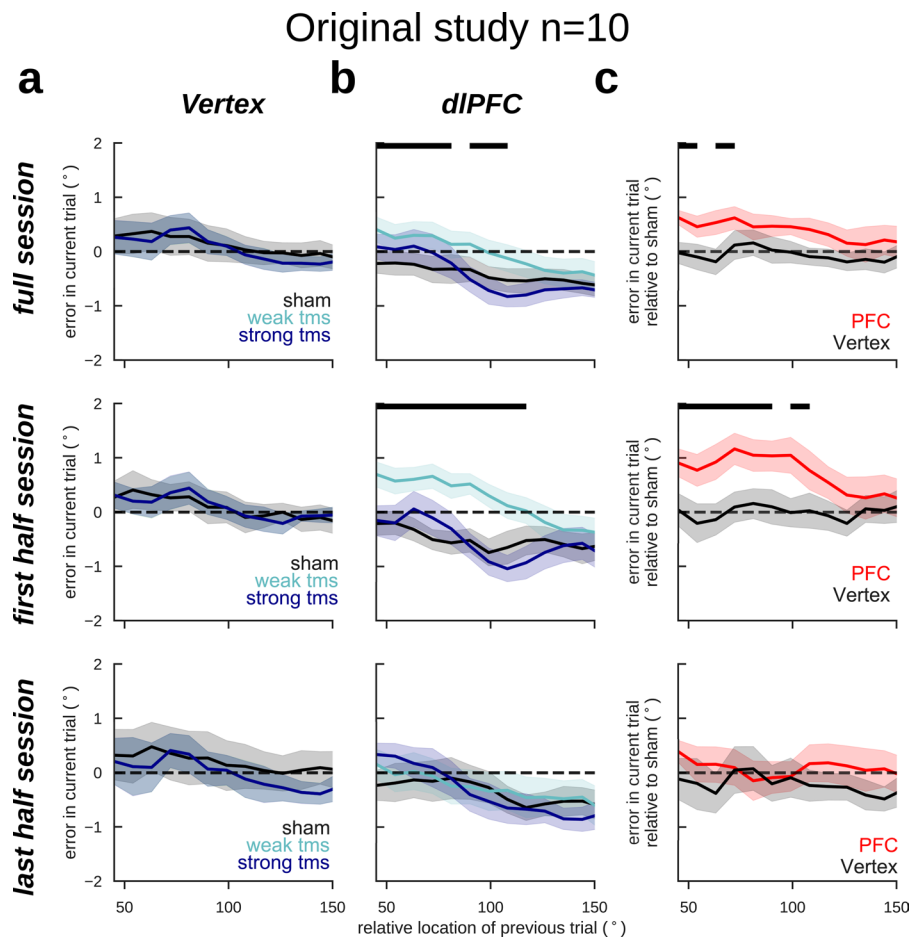
Extended Data Fig. 5 | Structured inhibition is necessary for repulsive serial biases at far distances. Top panel, illustration of two different models that have different inhibitory connectivity profiles. On the left, inhibitory connectivity strength from inhibitory to excitatory neurons is similar for all distances between their preferred locations. On the right, inhibition is structured such that similarly tuned neurons have stronger feedback inhibition. This shows that repulsive biases are caused by repulsive interactions between simultaneously active bumps in the network^{39,40}, and are absent when there is no reignited bump that recruits localized inhibition at the flanks of the pre-cue bump of activity.



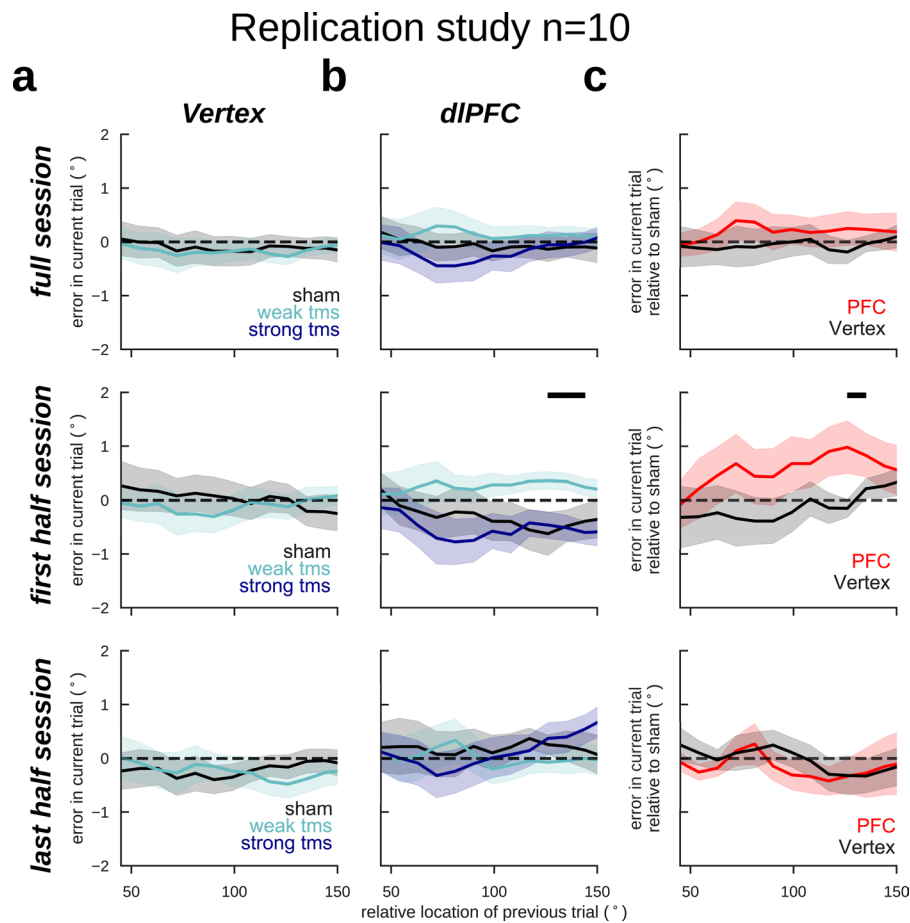
Extended Data Fig. 6 | Serial bias split between high-decoding and other trials (Fig. 5) is robust to the choice of different percentiles. **a**, In monkey behavior **b**, In human behavior. X-axis indicates quantiles used for the split in high- and low-decoding trials (Fig. 5), from a total of $n=1362$ trials in **a**, and a range of 792-908 trials per subject in **b**. Error bars are \pm s.e.m. (over $n=1362$ trials in **a**, and over $n=15$ subjects in **b**) and colored bars mark where corresponding difference in serial biases is different than zero ($p < 0.05$, two-sided bootstrap test).



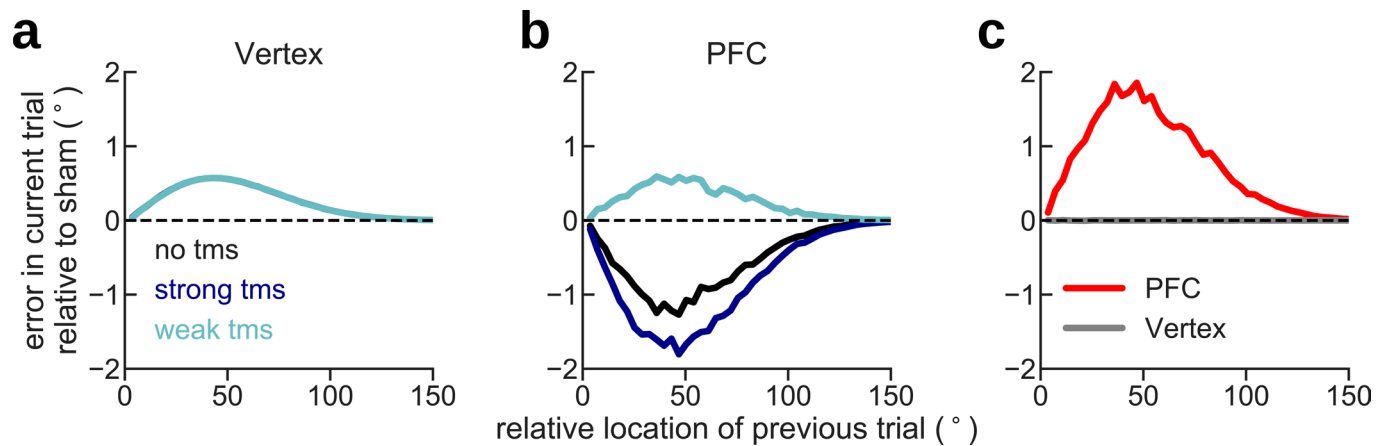
Extended Data Fig. 7 | The effect on serial biases of targeting dIPFC with TMS diminishes in the course of the experimental session. Serial bias plots averaged across $n=20$ independent subjects for trials with TMS applied in vertex (**a**) and PFC (**b**), and difference between serial biases computed for sham and weak-tms trials in vertex (black) and in PFC (red) blocks (**c**). Same analyses as in Fig. 6, but (top) analyzing trials from the full session, (middle) first half session (225 trials, replication of Fig. 6) and (bottom) last half session (225 trials). The behavioral impact of PFC TMS stimulation declined through the session, as if subjects desensitized ($prev-curr \times TMS\ intensity \times session-half$ $t_{11083} = -2.38$, $p = 0.017$. Methods, *Linear Mixed Models*). Serial biases were modulated by TMS in PFC, but not in Vertex ($prev-curr \times TMS\ intensity \times coil\ location$, $t_{18272} = 2.21$, $p = 0.027$. For dIPFC: $prev-curr \times TMS\ intensity$, $t_{11087} = 2.13$, $p = 0.032$. For Vertex: $t_{7166} = 0.03$, $p = 0.97$. Methods, *Linear mixed models*) when analyzing the full session, and analyzing only the first half session ($t_{9133} = 2.51$, $p = 0.011$). x-axis coordinates mark the central value of windows ($\pi/2$ radians, sliding by $\pi/30$ radians) used to calculate behavioral biases.



Extended Data Fig. 8 | Consistent fixation-period single-pulse TMS effects on serial biases: first experiment. Serial bias plots averaged across $n=20$ independent subjects for trials with TMS applied in vertex (**a**) and PFC (**b**), and difference between serial biases computed for sham and weak-tms trials in vertex (black) and in PFC (red) blocks (**c**). Same as Extended Data Fig. 6, but only analyzing data from the original study ($n=10$ subjects). Similarly to when pooling both the original and replication studies together, the behavioral impact of PFC TMS stimulation declined throughout the session, however not significantly ($prev-curr \times TMS\ intensity \times session-half$ $t_{5701} = -1.73$, $p = 0.08$, Methods, *Linear Mixed Models*). Serial biases were modulated by TMS in PFC, but not in Vertex ($t_{5705} = 1.92$, $p = 0.05$) when analyzing the full session, and analyzing only the first half session ($t_{3059} = 2.59$, $p = 0.009$, Methods). x-axis coordinates mark the central value of windows ($\pi/2$ radians, sliding by $\pi/30$ radians) used to calculate behavioral biases.



Extended Data Fig. 9 | Consistent fixation-period single-pulse TMS effects on serial biases: replication experiment. Serial bias plots averaged across $n=20$ independent subjects for trials with TMS applied in vertex (**a**) and PFC (**b**), and difference between serial biases computed for sham and weak-tms trials in vertex (black) and in PFC (red) blocks (**c**). Same as Extended Data Fig. 6 and 7, but only analyzing data from the pre-registered (<https://osf.io/rguzn/>) replication study ($n=10$ subjects). Similarly to the original experiment, the behavioral impact of PFC TMS stimulation declined throughout the session, however not significantly ($\text{prev-curr} \times \text{TMS intensity} \times \text{session-half}$ $t_{5375} = -1.63$, $p = 0.1$, Methods, *Linear Mixed Models*). Similarly to the original study, serial biases were more strongly modulated by TMS in PFC than in Vertex, however not significantly ($t_{5379} = 1.12$, $p = 0.25$) when analyzing the full session and the effect was stronger when analyzing only the first half-session ($t_{2675} = 1.91$, $p = 0.06$, Methods). x-axis coordinates mark the central value of windows ($\pi/2$ radians, sliding by $\pi/30$ radians) used to calculate behavioral biases.



Extended Data Fig. 10 | A phenomenological model of our hypothesis on how long-term physiological effects of single TMS pulses affect serial bias curves in event-related experimental sessions. Our TMS results show a difference between the effects of sham stimulation at the vertex and sham stimulation over dIPFC (Fig. 6). We interpret this baseline difference as the possible effect of long-term physiological alterations by single pulses 58 (but see ref. ⁷²) that carry over from “strong-tms” trials to “no-tms” trials. We explicitly implemented this interpretation in the following way: we generated trial-by-trial responses biased depending on the sequence of stimuli according to a given baseline serial bias curve (**a**, “Vertex” condition where TMS is ineffective). In the “PFC” condition the serial bias strength changed depending on TMS conditions: in “weak-tms” trials the pulse had the acute effect of increasing the bias strength momentarily by an additive factor (3 times the baseline bias strength), in “strong-tms” trials the effect of the pulse was chronic: the bias changed with a negative additive component (equal in magnitude to the baseline strength), which decayed slowly through subsequent trials (10% decay/trial). When collapsing together “responses” obtained on the basis of this model through a sequence of randomly selected “no-tms”, “weak-tms” and “strong-tms” trials, serial bias curves showed the pattern observed experimentally, where sham (“no-tms”) trials show repulsion in the “PFC” condition (panel **b**) and not in the “Vertex” condition (panel **a**). The difference of serial bias curves for “weak-tms” and “no-tms” then showed the modulation clearly in “PFC” and not in “Vertex” (panel **c**), as seen in the data (Fig. 6).

72. Romero, M. C., Davare, M., Armendariz, M. & Janssen, P. Neural effects of transcranial magnetic stimulation at the single-cell level. *Nat. Commun.* **10**, 2642 (2019).

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

EEG recordings used Deltamed Coherence Software version 5.1. Behavioral experiments with humans were programmed in Python 2.7 using Psychopy version 1.82.01.

Data analysis

Data were analyzed using custom scripts in Python 2.7 (monkey and TMS data) and in Python 3.7.4 (human EEG data). EEG data was pre-processed using Fieldtrip (version 20171231) in MATLAB R2017b and R2019a. The custom code used in this study is publicly available at <https://github.com/comptelab/interplayPFC>.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data that support the findings of this study are available at <https://github.com/comptelab/interplayPFC>.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For the analysis of monkey data we were not able to predetermine sample sizes because this was data acquired previously (Constantinidis et al 2001). For human data, sample sizes were based on relevant prior literature. In the case of the EEG study, we matched the sample size (n=15) to the one used in a previous study that successfully decoded memory contents from EEG in an identical task (Foster et al. 2015). In the case of the TMS study, we predetermined the sample size (n=10) considering that TMS-induced memory reactivations had been shown in a previous study with 6 participants (Rose et al. 2016). We validated the results in a replication experiment with the same sample size (n=10).
Data exclusions	<p>* No monkeys were excluded from the analysis. In the EEG study, one participant aborted because of physical discomfort. Another participant repeated the session on a different day because they aborted their first session with too few trial blocks. For this participant we only analyzed session 2. In the TMS study, one participant dropped the study when acquiring her MRI because she suspected pregnancy.</p> <p>* For neural data analyses, we excluded neurons without significant tuned delay activity. This was because of the hypothesis of our study (we wanted to explore the interaction between persistent and activity-silent mechanisms) and was predetermined in this study, as in other previous studies with this dataset (Constantinidis et al 2001; Compte et al. 2003; Wimmer et al. 2014).</p> <p>* For behavioral analyses, we excluded trials where behavioral reports were too far from the target to remove guess trials that may have not engaged working memory. For monkeys, this was done directly at acquisition time and could not be predetermined for this study (criterion report more than 20 degrees away from target). For humans, we excluded trials with responses further than 1 radian from targets in the angular direction and further than half the radius (2.25cm) in the radial direction.</p> <p>* For EEG analyses, we excluded outlier trials based on the voltage trace variance and alpha-power variance over each session. This is customary practice to remove EEG artifacts. Specific thresholds were set at the time of pre-processing of the data prior to final analyses.</p>
Replication	We designed a replication study for the TMS experiment, to test the bias-enhancing effects of weak TMS stimulation and the disappearance of the effects as the session progressed. The methods, hypotheses and even the analysis codes for this replication study were pre-registered (https://osf.io/rguzn) prior to acquiring the data. Methods were applied as literally pre-determined and the results were parallel to our previous findings, validating our results. In the manuscript we report the aggregated data (participants were independent for the 2 studies), as well as the individual data for each experiment (supplementary data).
Randomization	Our study had a within-subject design, so randomization of participants across groups is not relevant for the study. Conditions of interest were typically randomized in our design: cue locations were pseudo-randomly chosen in monkey studies, and both cue locations and delay lengths were random in human EEG studies. For TMS experiments, cue locations and TMS intensity were random during experimental blocks, and TMS coil location was kept constant in each block and alternated from block to block, the order being counterbalanced in the 2 sessions of the same participant.
Blinding	Blinding was not necessary in regard to participants because this was a within-subject design with randomized task contingencies. For the TMS study, the experimenter could not be blind to the location of the coil.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals Monkey subjects were four adult male rhesus macaques. Two of the animals were tested 20 years ago, when age reporting was

Laboratory animals	not customary. From their reported weights (Constantinidis et al. J. Neurosci. 21:3646, 2001) they were fully grown adults, so we can estimate the age at more than 6 years old. The ages of the other two animals reported in the study (with only behavioral data) were both 9 years old.
Wild animals	This study did not involve wild animals.
Field-collected samples	This study did not involve samples collected from the field.
Ethics oversight	All experiments were conducted in accordance with the guidelines set forth by the US National Institutes of Health, as reviewed and approved by the Yale University Institutional Animal Care and Use Committee, and by the Wake Forest University Institutional Animal Care and Use Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	We studied healthy controls. The study does not address any specific covariate of interest across individuals, but within-subject comparisons between trial types.
Recruitment	Participants were recruited from a volunteer database, mostly including people associated with the research institute and hospital, in all cases naïve to this study.
Ethics oversight	Research Ethics Committee of Hospital Clínic (Barcelona)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

A2 Supplementary material for Chapter 3.2

This section²² contains Supplementary material which has been published alongside the main article presented in Chapter 3.2.

²² This section has been published as:

Stein, H.*, Barbosa, J.*, Rosa-Justicia, M., Prades, L., Morató, A., Galan-Gadea, A., Ariño, H. Martinez-Hernandez, E., Castro-Fornieles, J., Dalmau, J. & Compte., A. (2020). Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia. *Nat Commun*, 11, 4250. <https://doi.org/10.1038/s41467-020-18033-3>. (*equal contribution)

Supplementary Material

Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia

Stein, Heike^{1,*}, Barbosa, Joao^{1,*}, Rosa-Justicia, Mireia^{1,2}, Prades, Laia¹, Morató, Alba¹, Galan-Gadea, Adrià¹, Ariño, Helena¹, Martinez-Hernandez, Eugenia^{1,3}, Castro-Fornieles, Josefina^{1,2,4}, Dalmau, Josep^{1,3,4,5,6}, Compte, Albert¹

¹ IDIBAPS, Barcelona, Spain

² Department of Child and Adolescent Psychiatry and Psychology, 2017SGR881, CIBERSAM, Institute Clinic of Neurosciences, Hospital Clínic, Barcelona, Spain

³ Service of Neurology, Hospital Clínic, Barcelona, Spain

⁴ Department of Medicine, University of Barcelona, Barcelona, Spain

⁵ Institució Catalana de Recerca i Estudis Avançats (ICREA)-IDIBAPS, Barcelona, Spain

⁶ Department of Neurology, University of Pennsylvania, USA

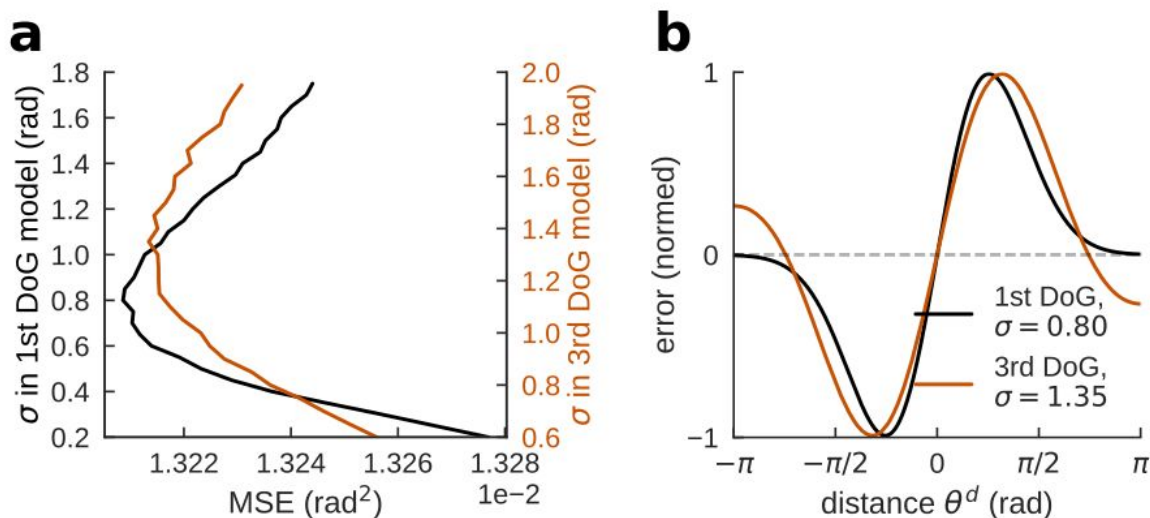
* These authors contributed equally

Corresponding author: acompte@clinic.cat

Supplementary Figures 1-15

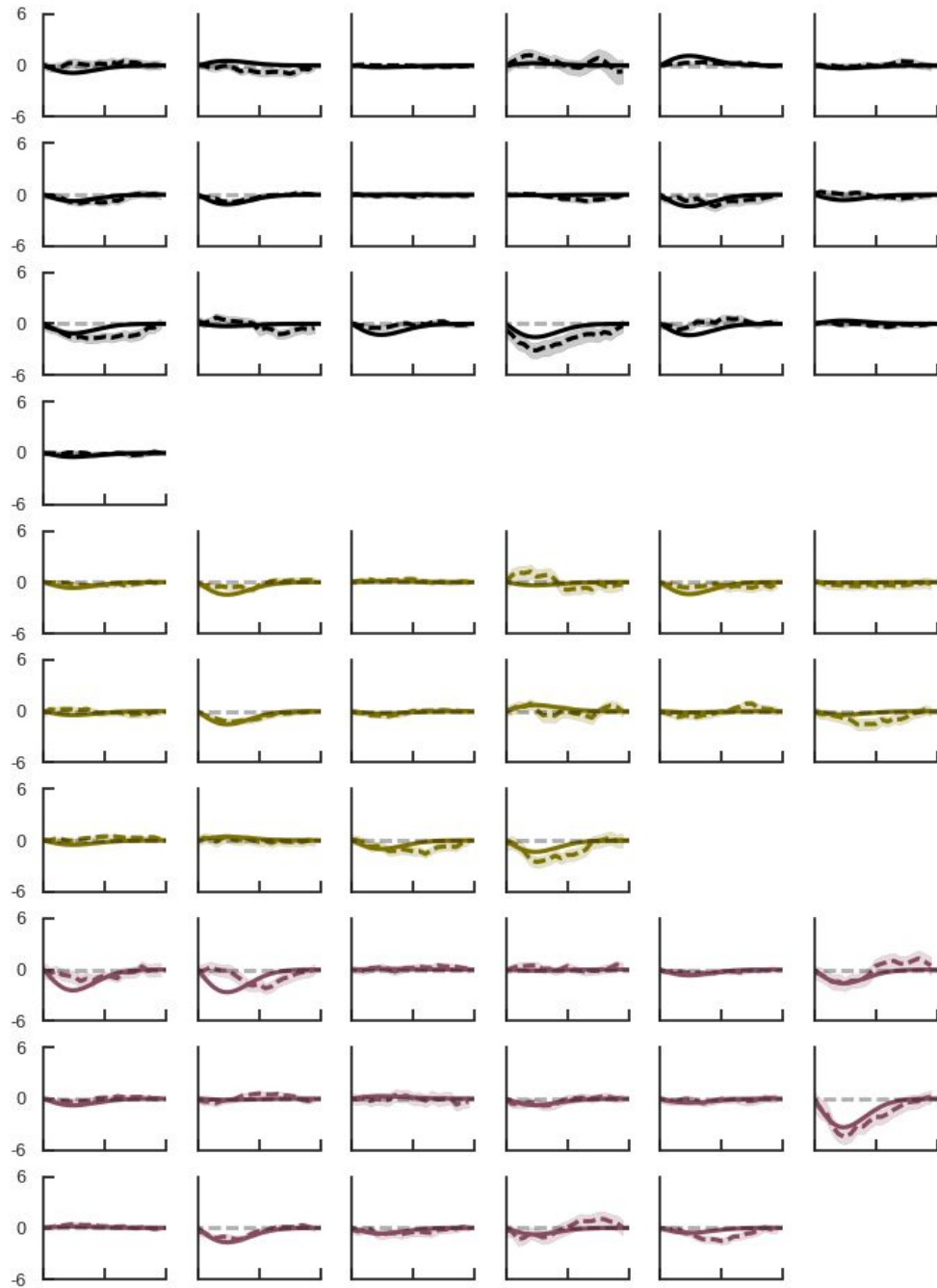
Supplementary Tables 1-2

Supplementary Figures



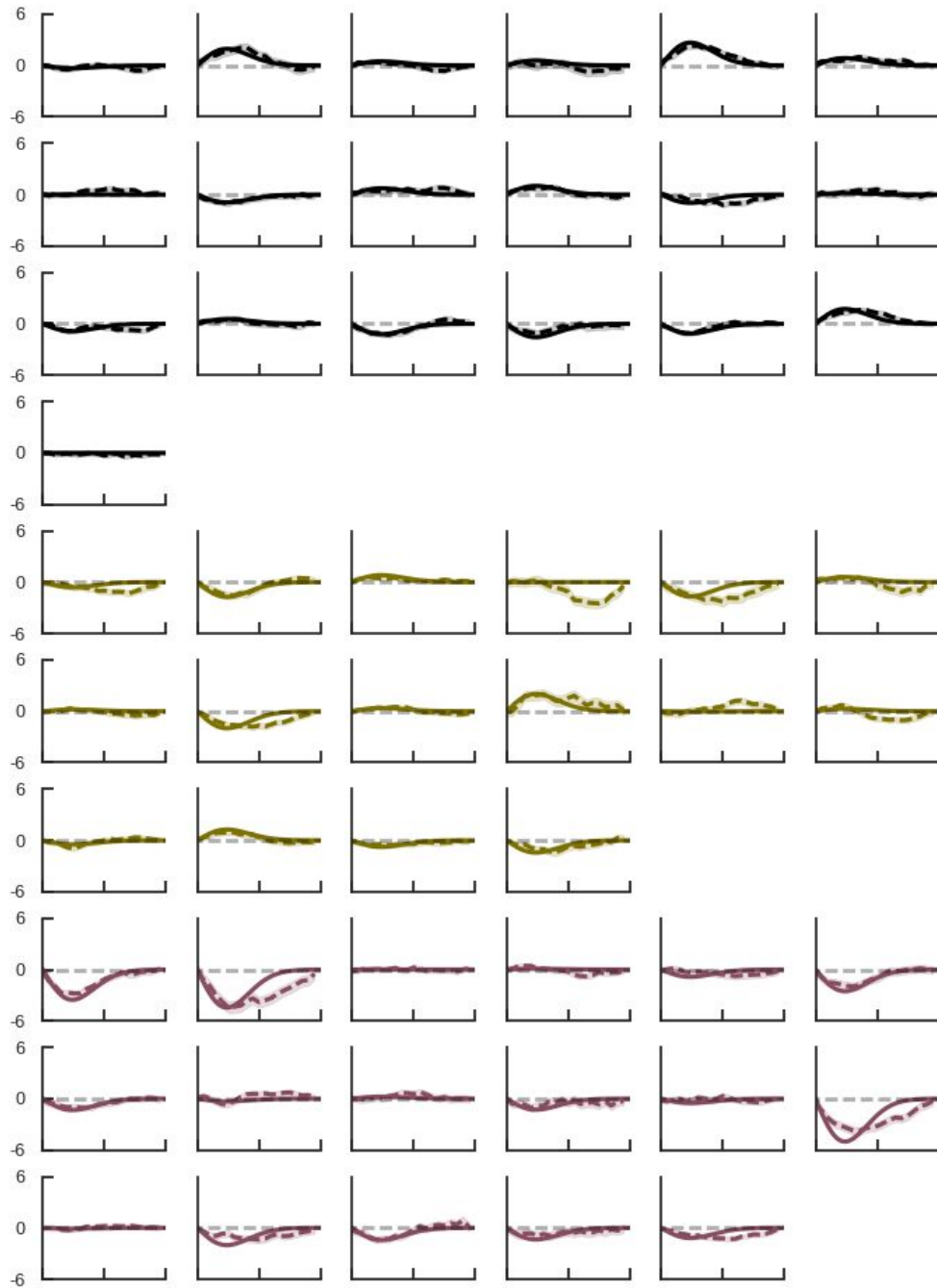
Supplementary Figure 1 | Hyperparameter cross-validation and model selection

a, Mean squared error for stratified hyperparameter optimization using cross-validation (1,000 repetitions, training set size = .33 from each subject) for first- (black) and third- (orange) derivative-of-Gaussian fits. Hyperparameters are different values of scale parameter σ of the underlying Gaussian with location hyperparameter $\mu = 0$. MSE: mean squared error. **b**, Shape of first- and third-derivative-of-Gaussian fits with optimal hyperparameter σ and $\mu = 0$. The cross-validation procedure used for model selection was carried out based on a model with a minimal set of variables (group, delay, and DoG(θ^d)), excluding random effects (Methods, equation (9)). Note that signed previous-current distances in radians were used in the linear model.



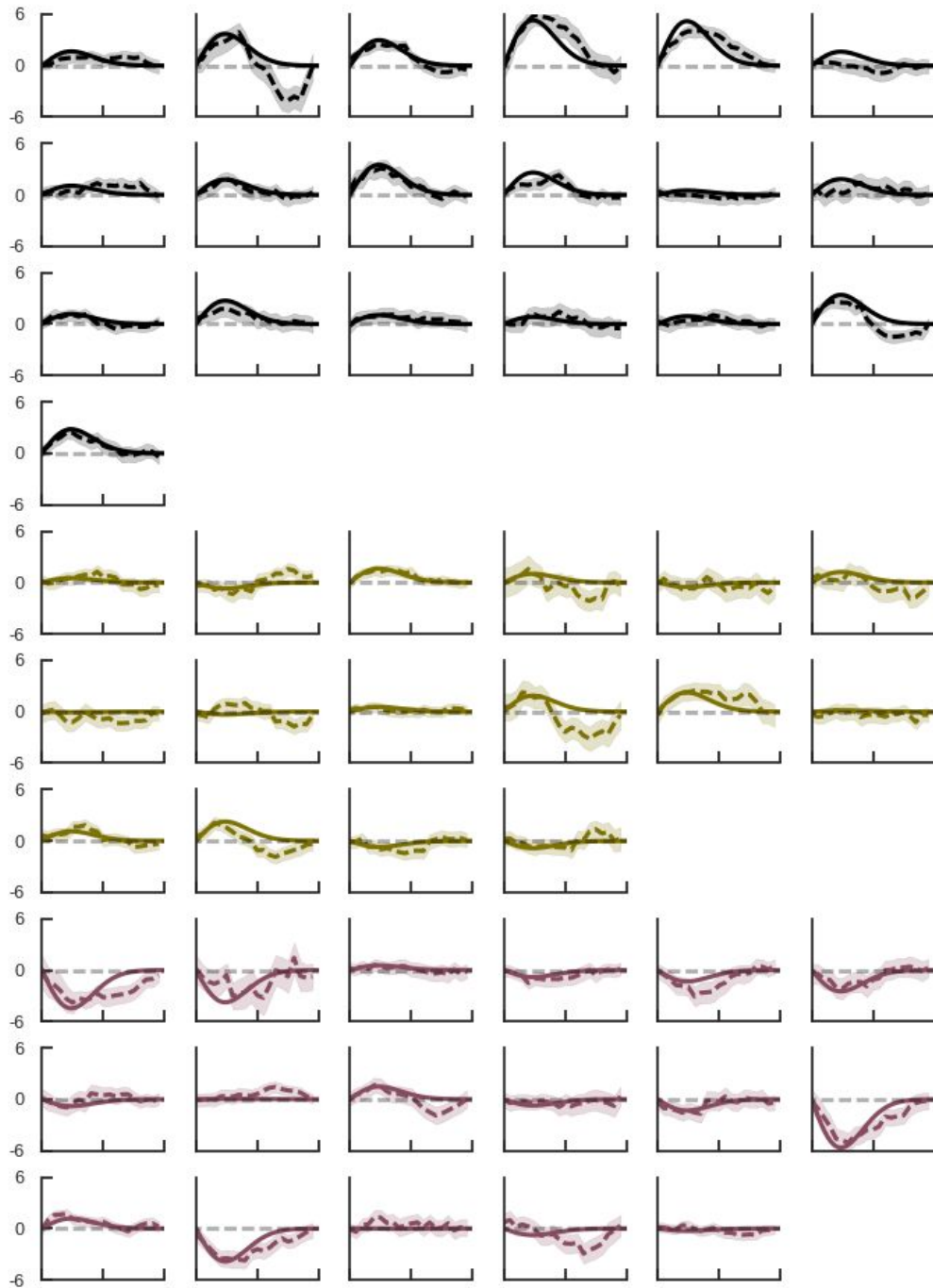
Supplementary Figure 2 | 0 seconds delay single subject bias and linear mixed model fit

Serial dependence is calculated for each subject as the ‘folded’ error θ^{e_s} (in degrees, y-axis) for different previous-current distances θ^d (x-axis, spanning absolute values of 0° - 180°) (dashed line; Methods). Shading, \pm s.e.m. Solid lines show linear model fits (Methods, equation (1)), omitting intercepts and negative values of θ^d for visualization. Black curves (row 1-4), ctrl, green curves (row 5-7), enc, purple curves (8-10), schz.



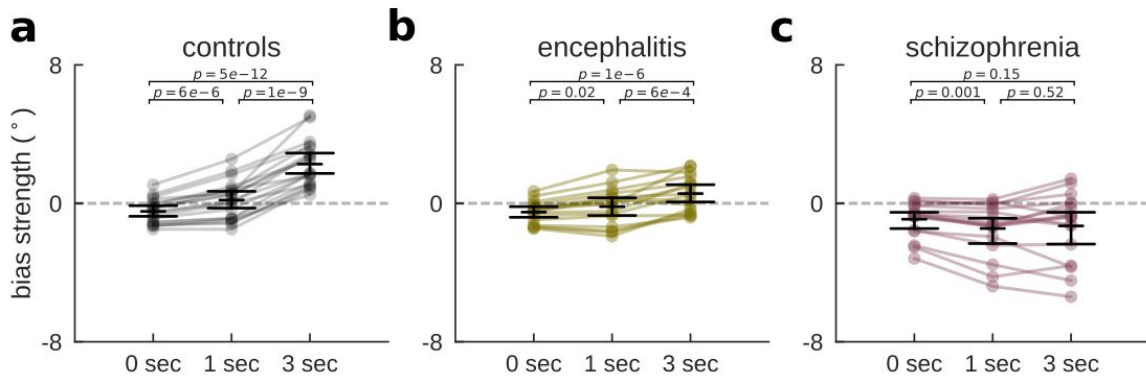
Supplementary Figure 3 | 1 second delay single subject bias and linear mixed model fit

Serial dependence is calculated for each subject as the ‘folded’ error θ^{st} (in degrees, y-axis) for different previous-current distances θ^{d} (x-axis, spanning absolute values of 0° - 180°) (dashed line; Methods). Shading, \pm s.e.m. Solid lines show linear model fits (Methods, equation (1)), omitting intercepts and negative values of θ^{d} for visualization. Black curves (row 1-4), ctrl, green curves (row 5-7), enc, purple curves (8-10), schz.



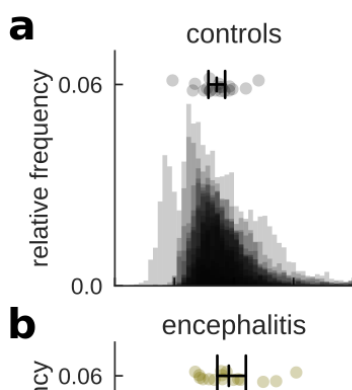
Supplementary Figure 4 | 3 seconds delay single subject bias and linear mixed model fit

Serial dependence is calculated for each subject as the ‘folded’ error θ^e (in degrees, y-axis) for different previous-current distances θ^d (x-axis, spanning absolute values of 0° - 180°) (dashed line; Methods). Shading, \pm s.e.m. Solid lines show linear model fits (Methods, equation (1)), omitting intercepts and negative values of θ^d for visualization. Black curves (row 1-4), ctrl, green curves (row 5-7), enc, purple curves (8-10), schz.



Supplementary Figure 5 | Serial dependence develops as a function of delay length

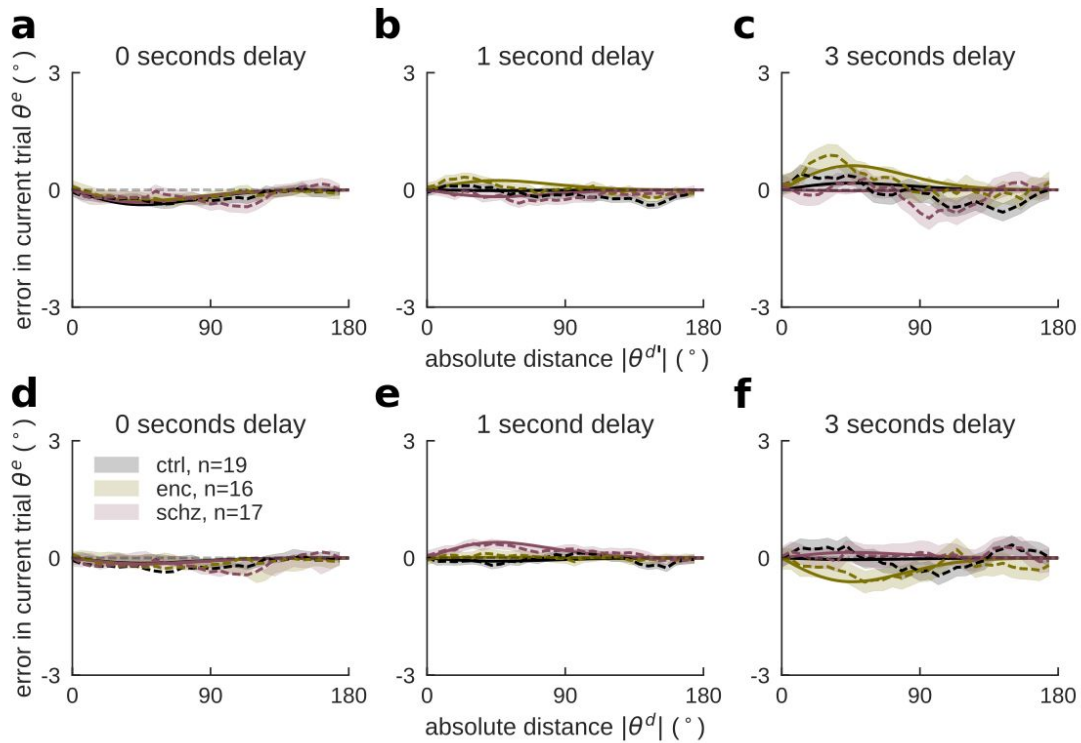
Individual (random coefficients; dots) and delay-specific group estimates (fixed effects; black horizontal lines indicate mean and bootstrapped 95% C.I. of the mean) of serial dependence. p-values report pairwise comparisons of random coefficients using paired, two-sided t-tests for $n = 19$ (ctrl), $n = 16$ (enc), and $n = 17$ (schz) patients. **a**, Initially repulsive biases became gradually more attractive with delay length for healthy controls (Methods, equation (3); delay \times DoG(θ^d), $F(2,17) = 26.91$, $p = 6e-6$; 0 vs 1 s: $t = -6.33$, $p = 6e-6$, Cohen's $d = -1.45$; 1 vs 3 s: $t = -11.37$, $p = 1e-9$, Cohen's $d = -2.6$; 0 vs 3 s: $t = -15.87$, $p = 5e-12$, Cohen's $d = -3.64$) and **b**, for encephalitis patients (delay \times DoG(θ^d), $F(2,23) = 5.06$, $p = 0.015$; 0 vs 1 s: $t = -2.71$, $p = 0.02$, Cohen's $d = -0.68$; 1 vs 3 s: $t = -4.32$, $p = 6e-4$, Cohen's $d = -1.08$; 0 vs 3 s: $t = -7.82$, $p = 1e-6$, Cohen's $d = -1.95$). **c**, schizophrenia patients' biases did not develop over the course of the delay (delay \times DoG(θ^d), $F(2,16) = 1.31$, $p = 0.30$; 0 vs 1 s: $t = 3.99$, $p = 0.001$, Cohen's $d = 0.97$; 1 vs 3 s: $t = -0.65$, $p = 0.52$, Cohen's $d = -0.16$; 0 vs 3 s: $t = 1.53$, $p = 0.15$, Cohen's $d = 0.37$), but stayed repulsive throughout all delay lengths (DoG(θ^d), $F(1,16) = 9.07$, $p = 0.008$).



Supplementary Figure 6 | Reduced serial dependence is not explained by group differences in ITI

Histograms of ITI lengths for **a**, control participants **b**, anti-NMDAR encephalitis, and **c**, schizophrenia patients. Here, the ITI is defined as the complete period from probe

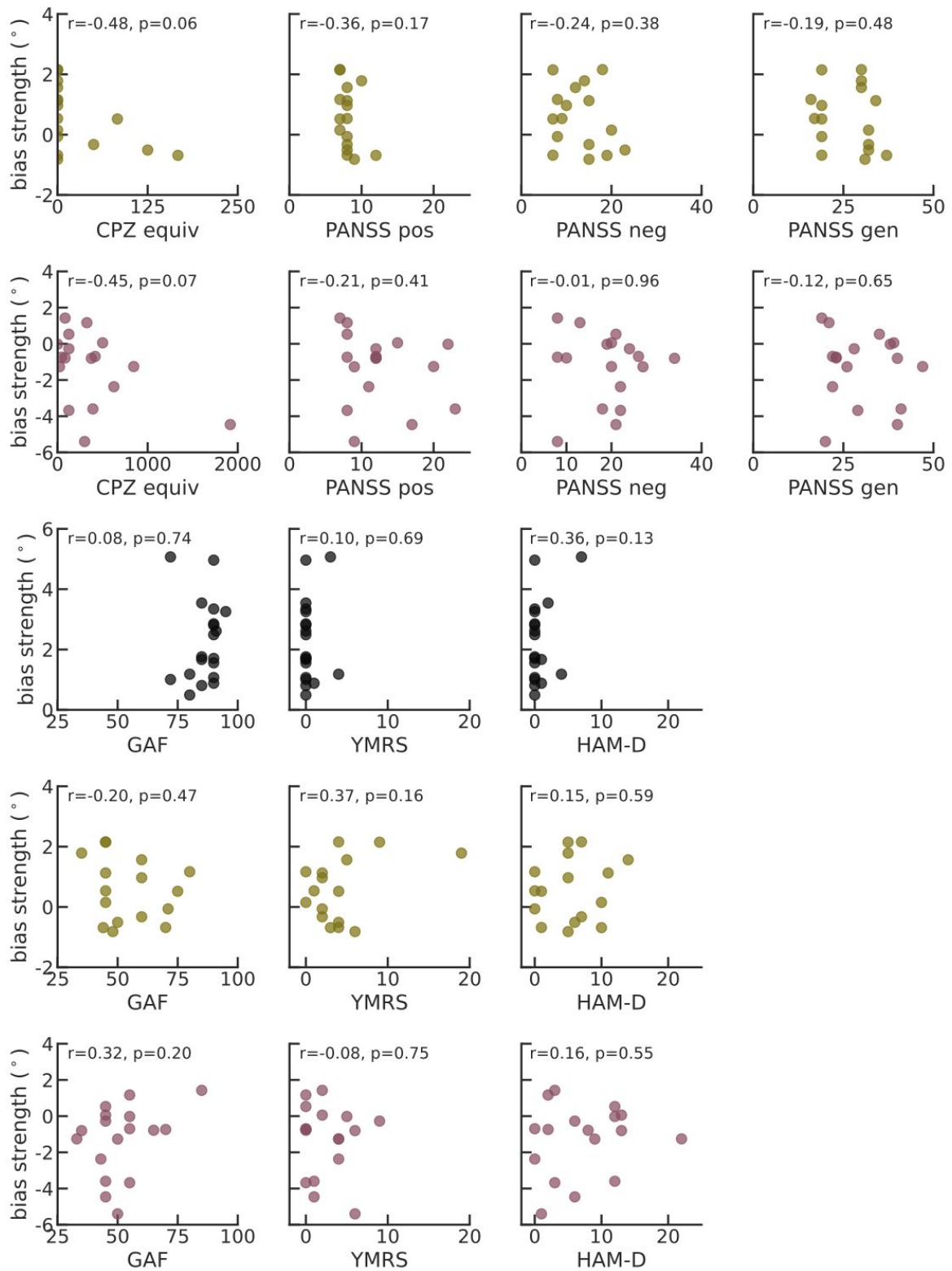
onset in trial n-1 to stimulus onset in trial n (including previous-trial response). Each plot shows normalized histograms, transparently overlaid for each participant. Points on top show median ITI lengths for each participant (n = 19 healthy controls, n = 16 patients with encephalitis, and n = 17 patients with schizophrenia), together with group mean and bootstrapped 95% C.I. (black middle line and error bars). There was a trend for longer median ITIs in patient groups (Kruskal-Wallis test for median ITI length, $H = 5.17$, $p = 0.08$; ctrl, 2.71 ± 0.32 s; enc, 2.91 ± 0.49 s; and schz, 3.03 ± 0.46 s; mean \pm s.d.). Including ITI \times DoG(θ^d) in our linear model (Methods, equation (4); $\Delta AIC = -13.8$) did not change group or delay effects of serial dependence (delay \times DoG(θ^d), $F(2,58) = 14.03$, $p = 1e-5$; group \times DoG(θ^d), $F(2,50) = 8.13$, $p = 9e-4$; group \times delay \times DoG(θ^d), $F(4,58) = 8.45$, $p = 2e-5$), but rather explained additional variance (ITI \times DoG(θ^d), $F(1,7503) = 15.76$, $p = 7e-5$).



Supplementary Figure 7 | Serial dependence to stimulus n-2 and stimulus n+1

a,b,c, An extended linear model with bias terms to both n-1 and n-2 stimuli (adding the θ^d -dependent term group \times delay \times DoG(θ^d), Methods, equation (7); Δ AIC = -4.11) showed significant delay-dependent bias towards the penultimate stimulus (delay \times DoG(θ^d), $F(2,52269) = 5.43$, $p=0.004$). Group differences could not be discarded (group \times DoG(θ^d), $F(2,52276) = 2.86$, $p = 0.06$), but there was no evidence for delay-dependent group differences (group \times delay \times DoG(θ^d), $F(4,52268) = 0.47$, $p = 0.76$). Groupwise models for each delay showed **a**, significant repulsive bias (DoG(θ^d), $F(1,8601) = 15.41$, $p = 9e-5$) but no group differences for delays of 0 s, (group \times DoG(θ^d), $F(2,8601) = 0.10$, $p = 0.91$). **b**, In contrast, groups differed for 1 s delays in absence of overall bias (DoG(θ^d), $F(1,34938) = 0.06$, $p = 0.81$; group \times DoG(θ^d), $F(2,34938) = 3.38$, $p = 0.03$), but **c**, not for 3 s delays (DoG(θ^d), $F(1,8669) = 2.57$, $p = 0.11$; group \times DoG(θ^d), $F(2,8669) = 1.55$, $p = 0.21$).

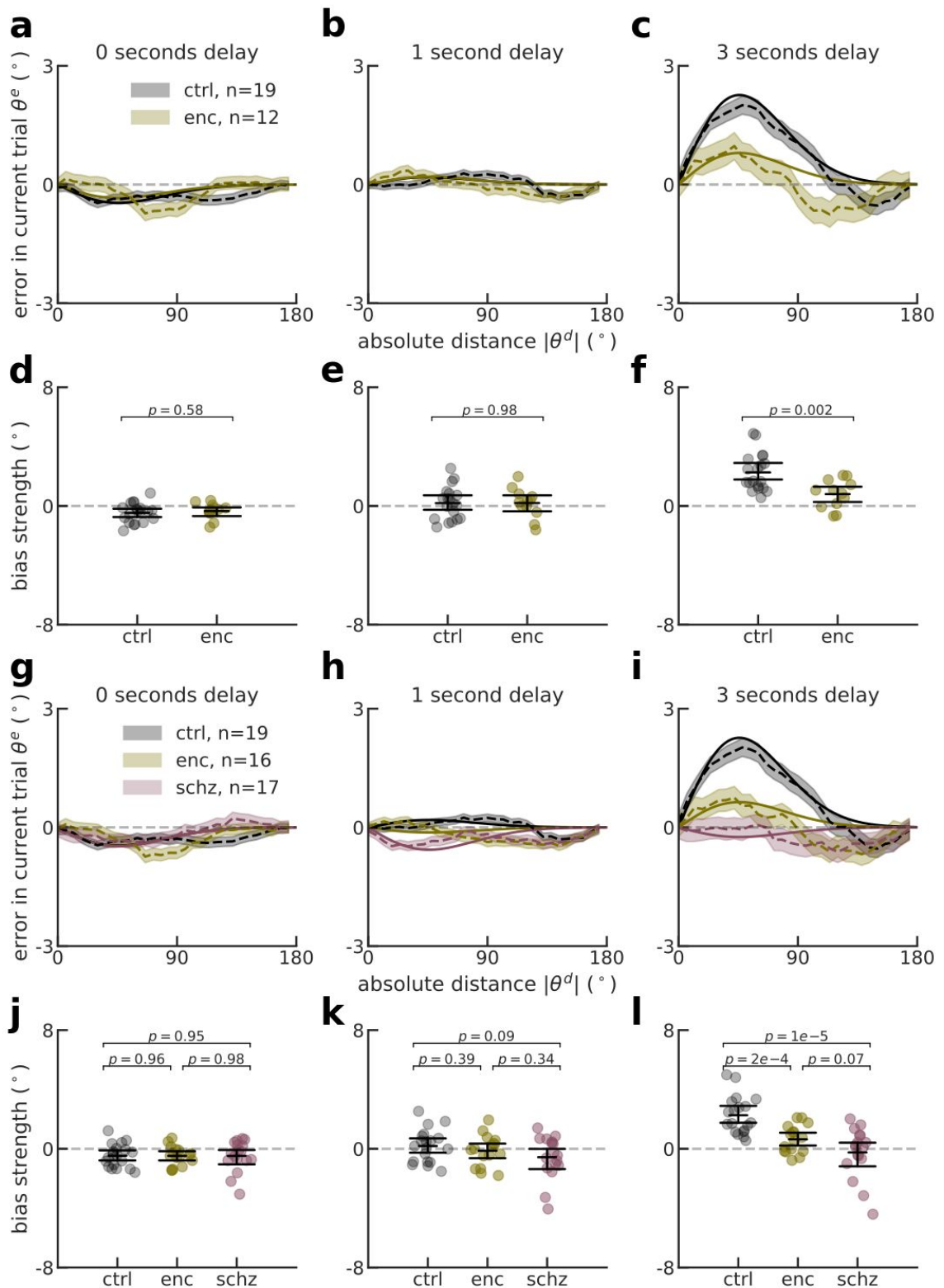
c, d, e, We investigated whether serial dependence to stimulus n-1 and group differences in biases could be explained by general response correlations. To detect potential spurious correlations across trials, we replaced previous-current distances (between trial n and trial n-1) in equation (1) with future-current distances (between trial n and trial n+1), as proposed in ref. ¹. There was no significant overall bias towards future stimuli (DoG(θ^d), $F(1,53) = 0.63$, $p = 0.43$; delay \times DoG(θ^d), $F(2,88) = 2.15$, $p = 0.12$; group \times DoG(θ^d), $F(2,53) = 1.57$, $p = 0.22$; group \times delay \times DoG(θ^d), $F(4,88) = 1.04$, $p = 0.39$), indicating non-significant contributions of general response correlations between trials to the reported group and delay effects of serial dependence. In all panels, dashed lines show ‘folded’ errors $\theta^{e'}$, and solid lines show linear model fits. Shading, \pm s.e.m. across pooled trials from $n = 19$ healthy controls (ctrl), $n = 17$ patients with schizophrenia (schz), and $n = 16$ patients with anti-NMDAR encephalitis (enc).



Supplementary Figure 8 | Correlations of serial dependence in 3 seconds delay trials with clinical scales

For each group, we correlated individual bias coefficients for 3 s delay trials (random effects) with clinical measures (Methods for description of administered tests). The strength of serial dependence did not correlate significantly with clinical scales. Correlations were

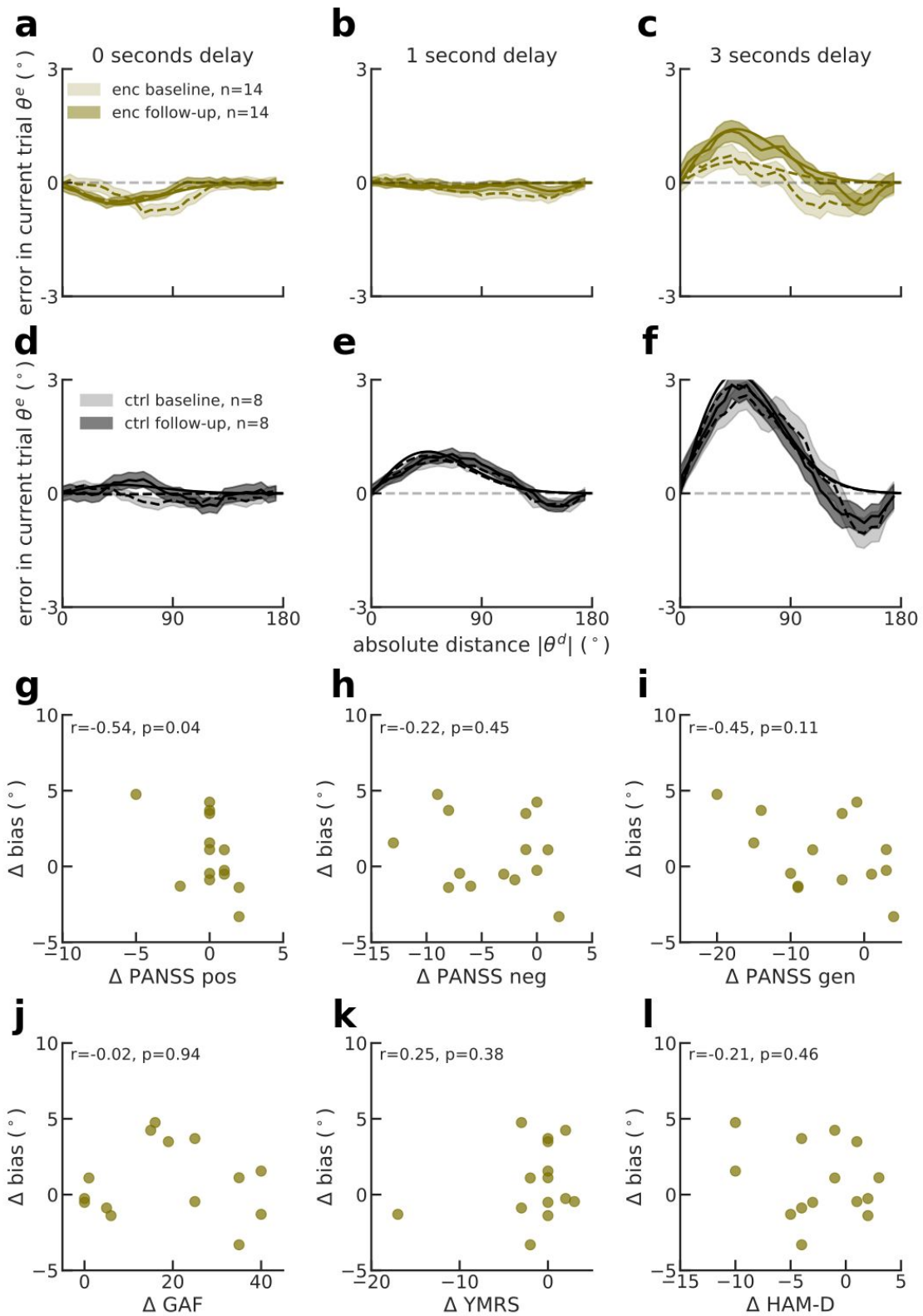
calculated using Pearson's r for $n = 19$ (ctrl, black, row 3), $n = 16$ (enc, green, rows 1 and 4), and $n = 17$ (schz, purple, rows 2 and 5). Correlations with antipsychotic medication (CPZ equivalent) reached marginal significance for both encephalitis and schizophrenia. To test whether medication could account for group differences in delay-dependent bias, we included a transversal estimate of antipsychotic medication, CPZ, as a covariate in our linear model (Methods, equation (5); $\Delta\text{AIC} = -2.7$). Antipsychotic medication explained a significant amount of variance in delay-dependent bias (CPZ \times delay \times DoG(θ^d), $F(3,60) = 3.06$, $p = .03$), but did not change the pattern of results (delay \times DoG(θ^d), $F(2,62) = 17.58$, $p = 9e-7$; group \times DoG(θ^d), $F(2,48) = 3.92$, $p = 0.03$; group \times delay \times DoG(θ^d), $F(4,62) = 4.43$, $p = 0.003$). Moreover, to be able to pool all subjects for each of the correlations, we modeled subject-wise bias strength in 3 s trials as a function of group and a second regressor, corresponding to each of the above clinical measures. Models with psychosis-related measures (CPZ and PANSS scales) were fitted on patient data only, and on all subjects' biases for all other models (including GAF, YMRS, HAM-D). In these analyses, only CPZ significantly predicted the strength of serial dependence (CPZ, $F(1,30) = 6.52$, $p = 0.02$), together with group (schz vs enc, $F(1,30) = 4.27$, $p = 0.05$). Measures: GAF (Global Assessment of Functioning Scale ²), YMRS (Young Mania Rating Scale ³), HAM-D (Hamilton Depression Rating Scale ⁴), PANSS (Positive and Negative Syndrome Scale ⁵) Positive, Negative and General Psychopathology Scale, CPZ equivalent (transversal estimate of antipsychotic medication as chlorpromazine equivalent).



Supplementary Figure 9 | Pronounced group differences in serial dependence remain after controlling for antipsychotic medication

a-f, To control for potential effects of chlorpromazine equivalent (CPZ, mg day⁻¹) on serial dependence, we fitted our full model as described in equation (1) on the unmedicated

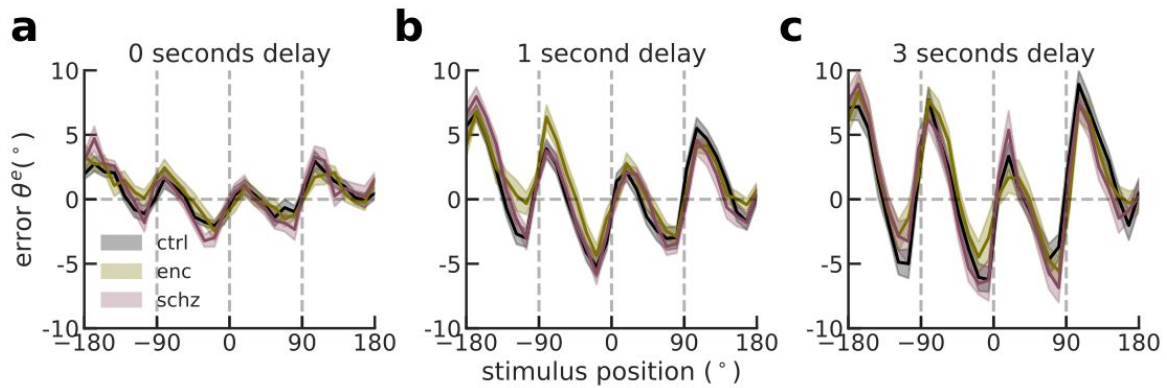
subset of participants (ctrl, $n = 19$ and enc, $n = 12$ out of $n = 16$, excluding schz due to $n = 1$). **a,b,c**, There were significant group differences in delay-dependent biases (group \times delay \times DoG(θ^d), $F(2,28) = 4.49$, $p = 0.02$ with $n = 12$, as compared to $F(2,32) = 5.70$, $p = 0.008$ with $n = 16$ enc, excluding the schizophrenia group), with no significant overall group differences in biases in neither the subset, nor the full group of patients (group \times DoG(θ^d), $F(1,29) = 1.29$, $p = 0.27$ for $n = 12$, and $F(1,33) = 3.91$, $p = 0.06$ with $n = 16$). **d,e,f**, Delay-wise models showed that the group difference in biases occurred in 3 s delay trials ($F(1,29) = 5.80$, $p = 0.02$ for $n = 12$), whereas biases were comparable between groups for shorter delays ($F(1,30) = 0.15$, $p = 0.70$ for 0 s, and $F(1,29) = 0.00$, $p = 0.99$ for 1 s). **g-l**, To obtain a conservative estimate of group differences after removing all possible linear effects of CPZ in all subjects ($n=19$ healthy controls, $n=16$ patients with encephalitis, and $n=17$ patients with schizophrenia), we first regressed trialwise errors on terms containing CPZ (equation (6)), and then estimated group and delay effects (as described in equation (1)) still present in the regression residuals from equation (6). CPZ had a significant effect on delay-independent and delay-dependent biases (CPZ \times DoG(θ^d), $F(1,52387) = 196.17$, $p < 1e-16$, and CPZ \times delay \times DoG(θ^d), $F(2,52387) = 4.91$, $p = 0.007$). **g,h,i**, Delay-independent and delay-dependent serial biases remained significantly altered in both patient groups after partially regressing CPZ equivalent from errors (group \times DoG(θ^d), $F(2,49) = 3.54$, $p = 0.04$, and group \times delay \times DoG(θ^d), $F(4,63) = 6.14$, $p = 0.0003$). **j,k,l**, Delay-wise models showed similar results to Figure 1, with equally repulsive serial dependence for all groups in 0 s trials (DoG(θ^d), $F(1,52) = 6.76$, $p = 0.01$, and group \times DoG(θ^d), $F(2,51) = 0.03$, $p = 0.97$) and group differences emerging in 3 s trials (group \times DoG(θ^d), $F(2,50) = 8.97$, $p = 0.0005$; with significant differences between individual estimates for ctrl vs enc, $t = 4.21$, $p = 2e-4$, Cohen's $d = 1.47$; ctrl vs schz, $t = 5.05$, $p = 1e-5$, Cohen's $d = 1.74$; enc vs schz, $t = 1.85$, $p = 0.07$, Cohen's $d = 0.66$), although not in 1 s trials (group \times DoG(θ^d), $F(2,48.1) = 1.38$, $p = 0.26$). **a-c, g-i**, Dashed lines with \pm s.e.m. shading, data; solid lines, linear model fits (Methods). ctrl: healthy controls, schz: schizophrenia, enc: anti-NMDAR encephalitis. **d-f, j-l**, Individual (random coefficients; dots) and group estimates of serial bias strength (fixed effects; black error bars indicate mean and bootstrapped 95% C.I. of the mean) by delay.



Supplementary Figure 10 | Serial dependence increases with encephalitis patients' recovery

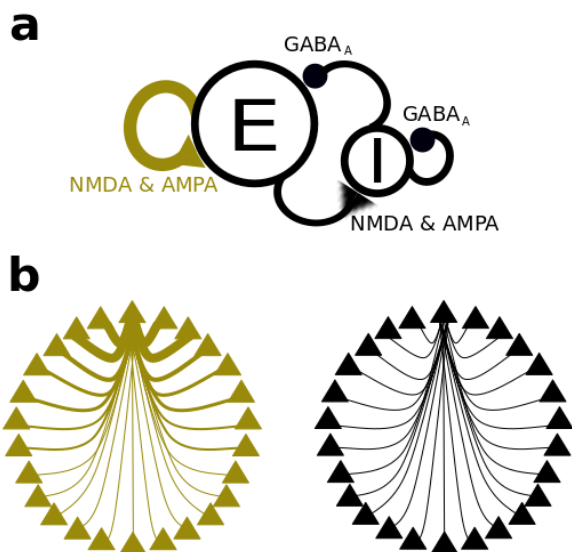
We performed a comparison of baseline and follow-up sessions (Methods, equation (8)) for $n=14$ encephalitis patients (enc, Supplementary Table 2) and $n=8$ controls (ctrl). Although the four-way interaction did not reach significance (session \times group \times delay \times DoG(θ^d),

$F(2,30124) = 0.79, p = 0.45$), group-wise models showed a normalization of biases in encephalitis patients' (**a-c**, session \times delay \times DoG(θ^d), $F(2,30124) = 3.07, p = 0.046$), and not in healthy controls (**d-f**, $F(2,16311) = 0.10, p = 0.90$). A delay-wise comparison of encephalitis patients' baseline and follow-up values showed that this difference was driven by biases in 3 s delays (session \times DoG(θ^d), $F(1,5030) = 4.43, p = 0.035$), while biases in 0 and 1 s delays did not change ($F(1,5030) = 0.15, p = 0.69$, and $F(1,20064) = 0.05, p = 0.81$, respectively). Note that due to the increased complexity of the model and the limited sample size, we could not model random effects in this model. In panels a-f, shading denotes mean \pm s.e.m. across pooled trials from all subjects of the respective group. **g-l**, To assess single-subject alterations in serial dependence and their correlation with clinical improvement, we estimated subject-wise models in 3 s delay trials for encephalitis patients, by modeling errors θ^e as a function of session, DoG(θ^d), and their interaction, session \times DoG(θ^d). We then correlated coefficients for session \times DoG(θ^d) (y-axis, Δ bias; positive values denote higher bias in the follow-up session) with change scores in clinical scales (x-axis; positive values denote higher scores in the follow-up session; Pearson's r and uncorrected p -values in panels indicate strength and significance of each correlation). We found that a more accentuated longitudinal reduction in PANSS positive symptoms was related to a stronger increase in memory-dependent biases ($r = -0.54$, C.I. = $[-0.83, -0.02]$, $p = 0.04$).



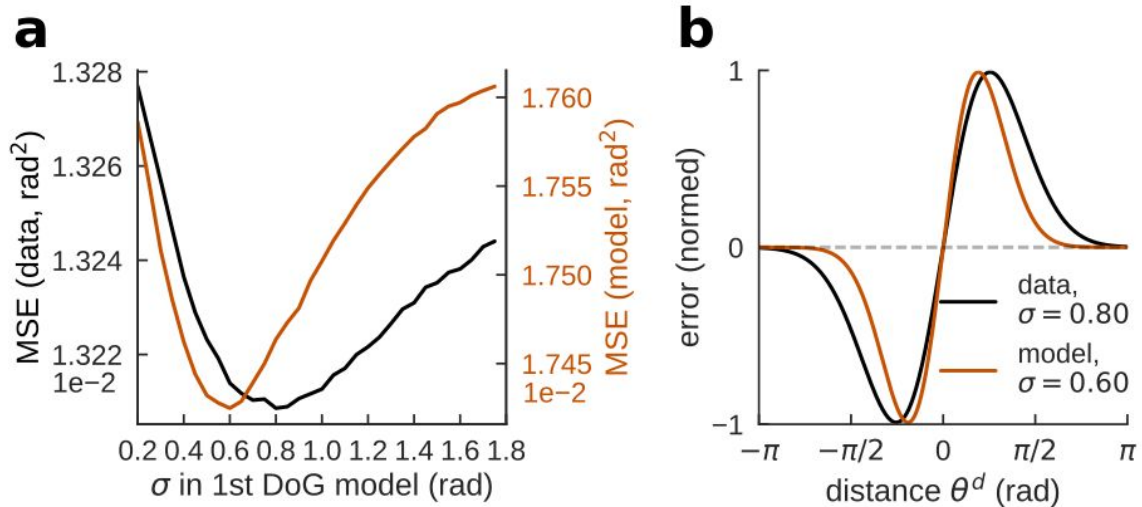
Supplementary Figure 11 | No group differences in working-memory biases with respect to cardinal directions

History-independent response biases, such as attractive or repulsive effects with respect to cardinal directions (0° , $\pm 90^\circ$, 180°) could influence behavioral performance and the fidelity of serial dependence estimations. **a,b,c**, Subject-averaged mean trialwise response errors θ^e , binned by stimulus location (x-axis, 30 bins of 12°) for each delay. This analysis revealed the effect of “repulsion from the axes” (e.g. ⁶⁻⁸) in all groups and delays. Error shading denotes mean \pm s.e.m. over $n=19$ healthy controls (ctrl), $n=17$ patients with schizophrenia (schz), and $n=16$ patients with anti-NMDAR encephalitis (enc). To quantify the strength of this effect, we measured the standard deviation (s.d.) of the binned statistic for each subject and delay, and assessed potential group- and delay-differences with an ANOVA. We observed stronger repulsion from the axes with increasing delays (observable in a,b,c; $F(2,147) = 72.45$, $p < 1e-16$), but no overall group differences ($F(2,147) = 1.72$, $p = 0.18$) or delay-dependent group differences ($F(4,147) = 0.16$, $p = 0.96$).



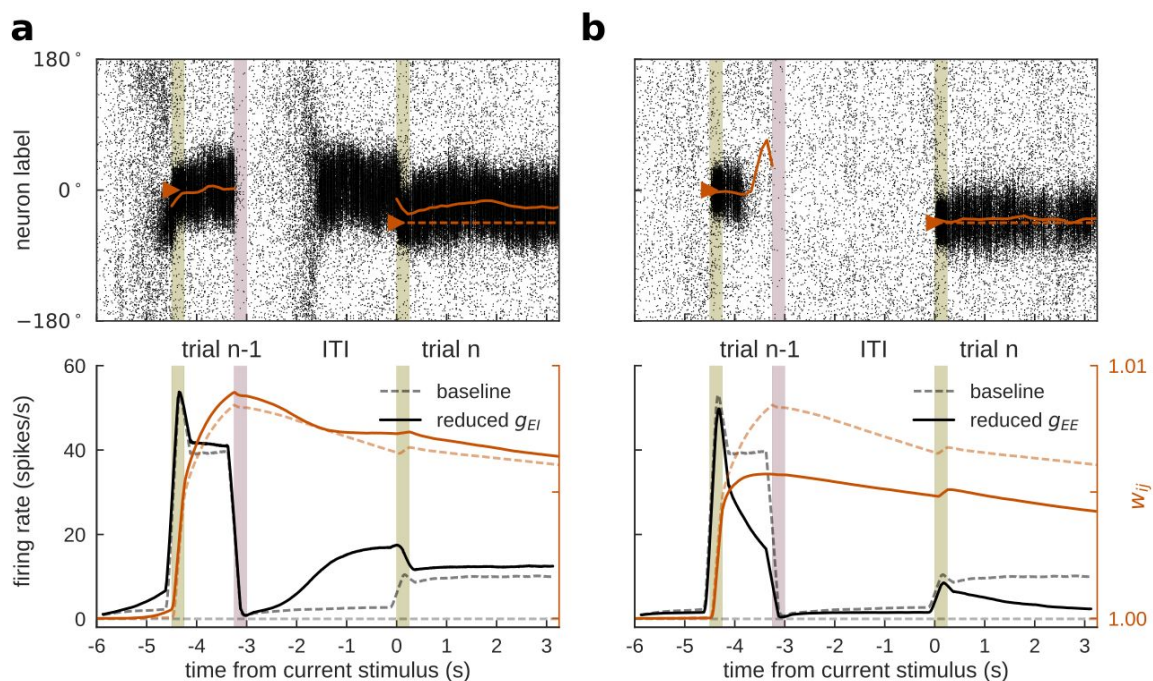
Supplementary Figure 12 | Network scheme and connectivity profile

a, Scheme of spiking neural network, consisting of 1,024 excitatory and 256 inhibitory neurons. Neurons from both pools were connected in an all-to-all fashion, with excitatory connections governed by NMDA and AMPA dynamics, and inhibitory connections governed by GABA_A dynamics. STP affected recurrent excitatory connections. **b**, Weight profiles for recurrent excitatory (green) and all other connections (black). For recurrent connections, weights between neurons preferring similar locations were higher, while more distant neurons were only weakly connected. All other connections had flat connectivity profiles, with equal weights between similar and dissimilar neurons.



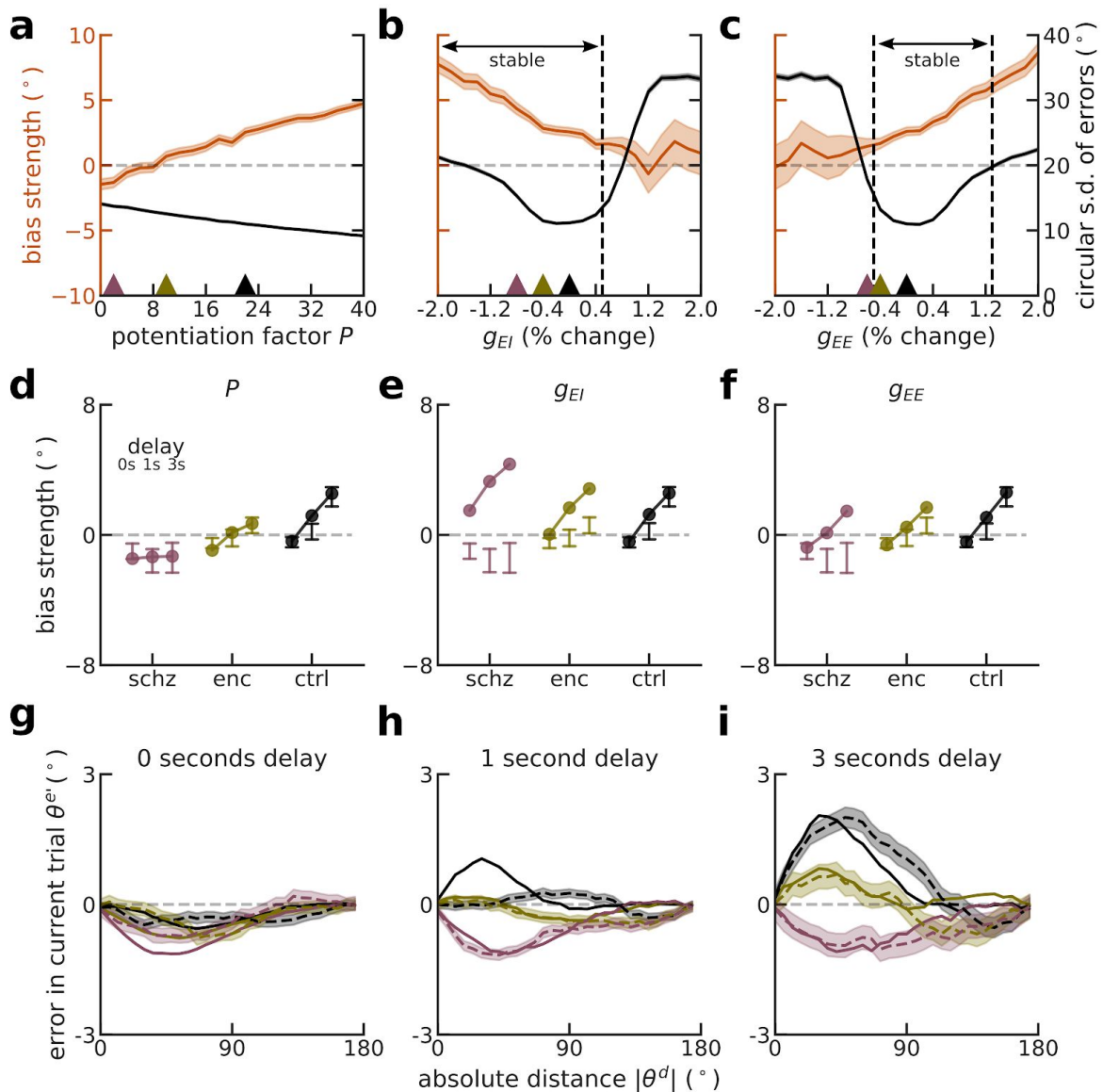
Supplementary Figure 13 | Hyperparameter cross-validation and model selection

a, Mean squared error for stratified hyperparameter optimization using cross-validation (1,000 repetitions, training set size = 0.33 from each subject) for data (black) and for the network model (orange; cross-validation with 1,000 repetitions, training set size = 0.33 of 21,000 simulated trials with baseline STP and conductance parameters, corresponding to the control condition in Figure 3). Hyperparameters are different values of scale parameter σ (in radians) of the underlying Gaussian with location hyperparameter $\mu = 0$. **b**, Shape of first-derivative-of-Gaussian fits with optimal hyperparameter σ and $\mu = 0$ for data (black) and model (orange). Hyperparameter cross-validation for neural network simulations was carried out for the default parameter values of P , g_{EI} and g_{EE} as reported in Methods. Note that in **b**, signed previous-current distances are indicated in radians as used in the linear model.



Supplementary Figure 14 | Network behavior for reduced g_{EI} and g_{EE}

a, For reduced g_{EI} , network activity was disinhibited and baseline firing became unstable. Spontaneous activity bumps emerged in the ITI (upper panel), often in neurons that had been active during the previous delay. Lower panel shows firing rates and STP traces at neurons selective to stimuli appearing at 0° for the baseline condition (0% reduction, dashed lines) and the disinhibited condition (1% reduction, solid lines), averaged over 1,000 trials in which the second stimulus appeared at randomized locations. **b**, For reduced g_{EE} , delay firing became unstable and active working memory representations were lost over the delay (upper panel). Lower panel analogous to **a**, for the baseline condition (0% reduction, dashed lines) and the condition of reduced excitation (1% reduction, solid lines). Lower panels were computed as in Figure 2 but including trials for which the same neurons were coactive in the two successive trials. This explains the difference between dashed lines here and in Figure 2b, trial n.



Supplementary Figure 15 | Modeling results are generalizable to networks with STP in both E-E and E-I synapses

Network simulations as in Figure 3, but for STP (and reductions in STP for **a,d,g-i**) in both recurrent excitatory and excitatory-to-inhibitory synapses. **a,b,c**, Serial dependence (orange, bias coefficients from linear model, Methods) and precision (black, circular s.d. of errors) as a function of model parameters in 3 s delay trials (20,000 trials per parameter value). Vertical dashed lines indicate transition to ‘unstable’ network regimes for which more than 10% of trials were outliers ($|\theta^e| > 57.3^{\circ}$, i.e. 1 radian). Shading, 95% C.I. for parameter estimates. **a**, Serial dependence decreased gradually when decreasing STP (potentiation factor P), while the network remained stable for all simulated values of P . Precision changed slightly as a function of STP. **b**, Cortical disinhibition via decreased g_{EI} augmented serial bias while strongly affecting precision and stability. Note that for the observed parameter values, the network did not enter the disinhibited unstable state (left), unlike in Figure 3. **c**, Lowering recurrent cortical excitation (g_{EE}) led to the opposite pattern, decreasing biases. **d,e,f**, delay dependence of biases for each group, as defined by

parameter values in **a,b,c**, (respectively colored triangles). Points depict mean bias strength (over 20,000 trials) for each parameter value. For comparison, error bars indicate 95% CI for bias strength obtained from $n = 19$ healthy controls (ctrl), $n = 17$ patients with schizophrenia (schz), and $n = 16$ patients with anti-NMDAR encephalitis (enc) (reordered from Figure 1g-i). **d**, Lowering STP strength reproduced the experimental data. In **e** and **f**, reduction of NMDAR conductances (g_{Ei} or g_{EE}) did not reproduce group and delay dependencies of experimental biases. **g,h,i**, Solid lines, simulated serial dependence by delay length for different values of P , indicated by colored triangles in **a** (20,000 trials per potentiation level P). Dashed lines with error bars, serial dependence in encephalitis, schizophrenia, and healthy controls. Bias calculated as averaged ‘folded’ error θ^{es} for binned absolute previous-current distances θ^d . Shading, \pm s.e.m. Compare to Figure 3 in the main text for a network with STP (and STP disruptions in patients) only in E-E connections.

Supplementary Tables

	ctrl (n=19) mean (s.d.) vs enc/schz	enc (n=16) mean (s.d.) vs ctrl/schz	schz (n=17) mean (s.d.) vs ctrl/enc	F-value / Chi-square	p-value (two-tailed)
age (years)	22.4 (6.8)	25.5 (6.6)	20.2 (6.1)	2.71	0.08
gender (% male)	21.1	12.5	41.2	3.86	0.14
medication (% taking antipsychotics)	0.0 n.s. / *	25.0 n.s. / *	93.3 * / *	35.35	2e-8
medication (CPZ equivalent, mg day ⁻¹)	0.0 (0.0) n.s. / *	26.6 (52.7) n.s. / *	370.6 (462.4) * / *	10.44	2e-4
Global Assessment of Functioning Scale	86.3 (6.3) * / *	54.9 (13.4) * / n.s.	51.5 (12.7) * / n.s.	54.62	3e-13
PANSS Positive Symptoms	7.2 (0.4) n.s. / *	8.1 (1.3) n.s. / *	12.5 (5.1) * / *	15.32	7e-6
PANSS Negative Symptoms	7.0 (0.0) * / *	12.9 (5.2) * / *	18.9 (7.4) * / *	24.29	5e-8
PANSS General Psychopathology	16.5 (1.2) * / *	26.0 (7.3) * / n.s.	30.2 (9.1) * / n.s.	20.40	4e-7
PANSS Total Score	30.7 (1.6) * / *	47.1 (12.7) * / *	61.6 (17.5) * / *	28.70	6e-9
Hamilton Depression Rating Scale	0.8 (1.8) * / *	5.4 (4.3) * / n.s.	7.3 (6.1) * / n.s.	10.70	1e-4
Young Mania Rating Scale	0.4 (1.1) * / n.s.	4.2 (4.6) * / n.s.	2.6 (2.8) n.s. / n.s.	6.72	0.003

Supplementary Table 1 | Clinical and demographic statistics of the population

Measures: GAF (Global Assessment of Functioning Scale ²), YMRS (Young Mania Rating Scale ³), HAM-D (Hamilton Depression Rating Scale ⁴), PANSS (Positive and Negative Syndrome Scale ⁵) Positive, Negative and General Psychopathology Scale, CPZ equivalent (transversal estimate of antipsychotic medication as chlorpromazine equivalent). The significance of pairwise post-hoc Tukey/Bonferroni-corrected Chi-square tests is reported below group mean and s.d. (n.s. marks non-significant comparisons, and * significant comparisons with FWE = 0.05).

	baseline (n=14) mean (s.d.)	follow-up (n=14) mean (s.d.)	t-value / Chi-square	p-value (two-tailed)
--	---------------------------------------	--	---------------------------------	--------------------------------

medication (% taking antipsychotics)	21.4	7.1	0.29	0.59
medication (CPZ equivalent, mg day ⁻¹)	21.4 (48.7)	3.0 (11.2)	1.34	0.20
Global Assessment of Functioning Scale	53.4 (12.4)	72.1 (13.7)	-4.69***	4e-4
PANSS Positive Symptoms	8.2 (1.4)	8.2 (0.9)	0.00	1.00
PANSS Negative Symptoms	12.6 (4.6)	8.6 (2.4)	3.24**	0.006
PANSS General Psychopathology	26.3 (7.1)	20.6 (3.2)	2.86*	0.01
PANSS Total Score	47.1 (12.1)	37.4 (5.5)	2.85*	0.01
Hamilton Depression Rating Scale	5.8 (4.4)	3.4 (4.0)	2.12	0.05
Young Mania Rating Scale	4.5 (4.8)	3.1 (2.3)	1.11	0.29

Supplementary Table 2 | Baseline/follow-up comparison of anti-NMDAR encephalitis patients

Measures: GAF (Global Assessment of Functioning Scale ²), YMRS (Young Mania Rating Scale ³), HAM-D (Hamilton Depression Rating Scale ⁴), PANSS (Positive and Negative Syndrome Scale ⁵) Positive, Negative and General Psychopathology Scale, CPZ equivalent (transversal estimate of antipsychotic medication as chlorpromazine equivalent).

References

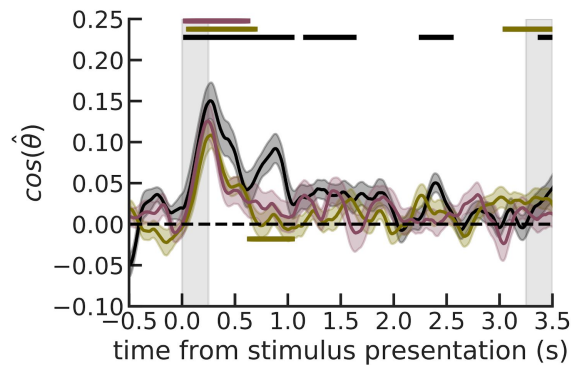
1. Cicchini, G. M., Anobile, G. & Burr, D. C. Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform. *Proc Natl Acad Sci USA* **111**, 7867–7872 (2014).
2. Bobes, J., Portilla, M. P. G., Bascarán, M. T., Saiz, P. A. & Bousoño, M. *Banco de instrumentos básicos para la práctica de la psiquiatría clínica*. (Psiquiatría Editores S.L., 2004).
3. Colom, F. *et al.* [Spanish version of a scale for the assessment of mania: validity and reliability of the Young Mania Rating Scale]. *Med Clin (Barc)* **119**, 366–371 (2002).
4. Ramos-Brieva, J. A. & Cordero-Villafafila, A. A new validation of the Hamilton Rating Scale for Depression. *J. Psychiatr. Res.* **22**, 21–28 (1988).
5. Kay, S. R., Fiszbein, A., Vital-Herne, M. & Fuentes, L. S. The Positive and Negative Syndrome Scale--Spanish adaptation. *J. Nerv. Ment. Dis.* **178**, 510–517 (1990).
6. Shin, H., Zou, Q. & Ma, W. J. The effects of delay duration on visual working memory for orientation. *J. Vis.* **17**, 10 (2017).
7. Wei, X.-X. & Stocker, A. A. A Bayesian observer model constrained by efficient coding can explain “anti-Bayesian” percepts. *Nat. Neurosci.* **18**, 1509–1517 (2015).
8. Lipinski, J., Simmering, V. R., Johnson, J. S. & Spencer, J. P. The role of experience in location estimation: Target distributions shift location memory biases. *Cognition* **115**, 147–153 (2010).

A3 Supplementary material for Chapter 3.3

This section²³ contains Supplementary material for the main manuscript presented in Chapter 3.2.

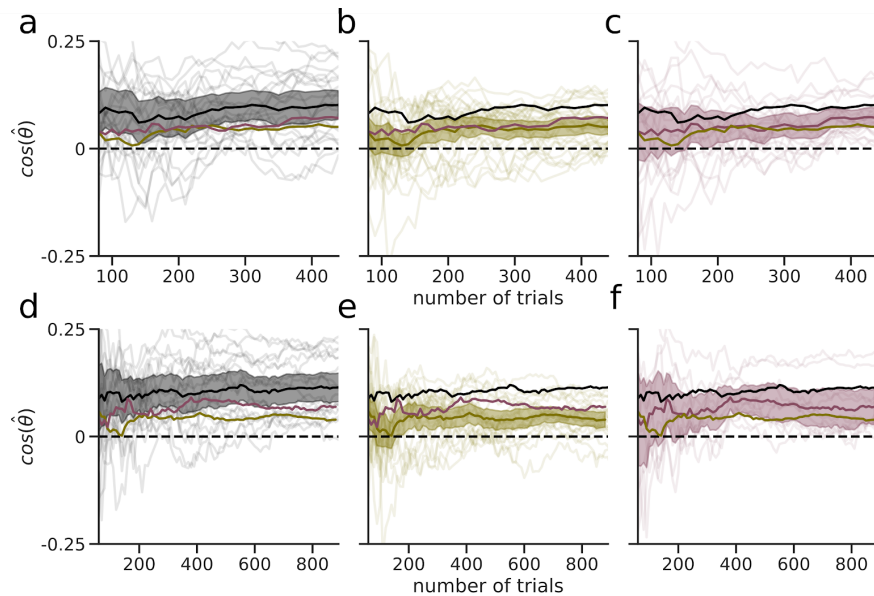
²³ This section is the Supplementary Material of a manuscript in preparation.

Supplementary Material



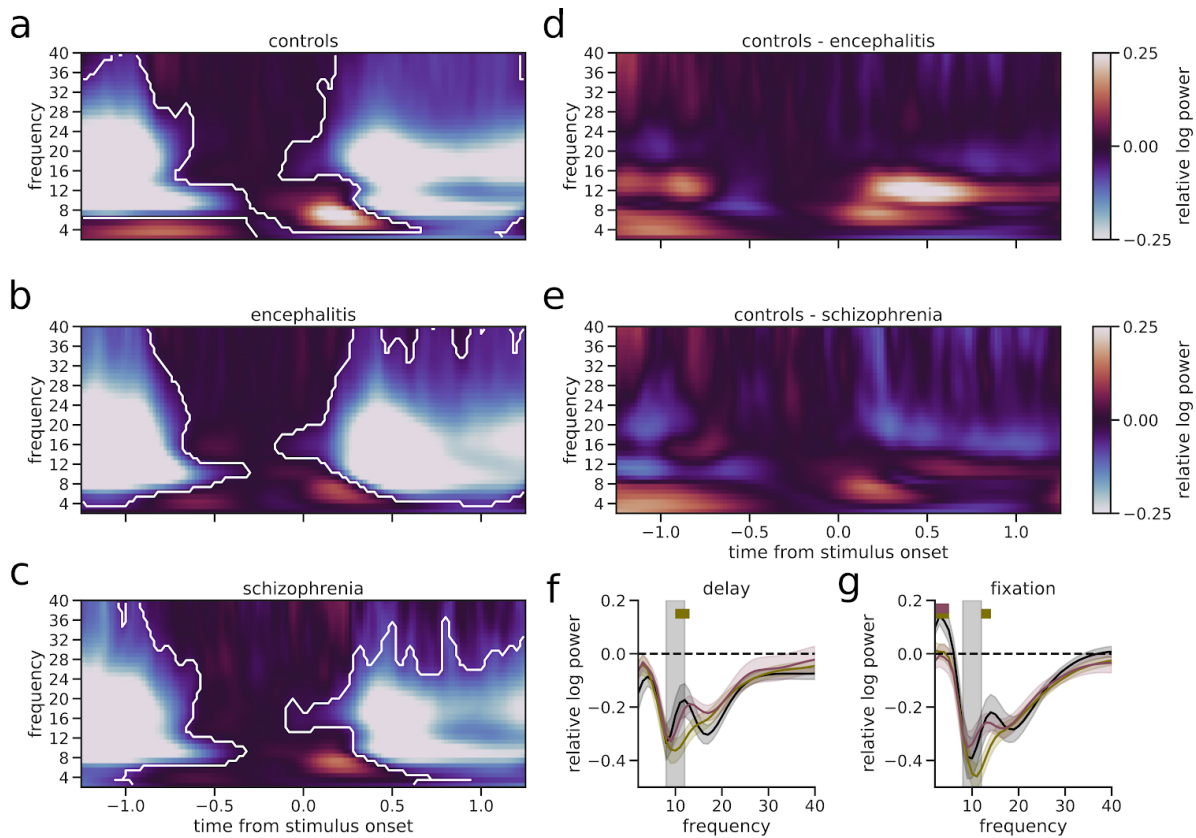
Supplementary Fig. 1 | Working memory codes fade during 3 s delays in all groups

Spatial locations in working memory decoded in 3 s memory delay trials in all groups. Compare to Fig. 3b, where the decoder was trained and tested on pooled 1 s and 3 s trials ($n = 900.63 \pm 103.02$ trials for ctrl, $n = 870.44 \pm 141.93$ for enc, and $n = 833.16 \pm 153.33$ for schz, mean \pm s.d.), leading to an increase in decoding performance due to decreased noise. Here, only the subset of 3 s delay trials was used for training and testing ($n = 175.82 \pm 20.61$ trials for ctrl, $n = 170.85 \pm 29.07$ for enc, and $n = 164.44 \pm 30.70$ trials for schz, mean \pm s.d.). Hence, decoding performance was reduced and more noisy. In ctrl, memory contents could be decoded until ~ 1.5 s after stimulus onset, while in enc and schz, codes already disappeared after ~ 0.75 s (upper significance bars, 1-sample cluster permutation test at $\alpha = 0.05$). Significant differences between groups (ctrl vs. enc) were only found for a period between ~ 0.6 s - 1 s (lower significance bars, 2-sample cluster permutation test at $\alpha = 0.05$), where healthy controls' code increased suddenly. Shading, 32% bootstrap C.I. of the mean across participants for $n = 22$ healthy controls (ctrl), $n = 19$ patients with schizophrenia (schz), and $n = 27$ patients with anti-NMDAR encephalitis (enc). Grey bars indicate stimulus and probe presentation.



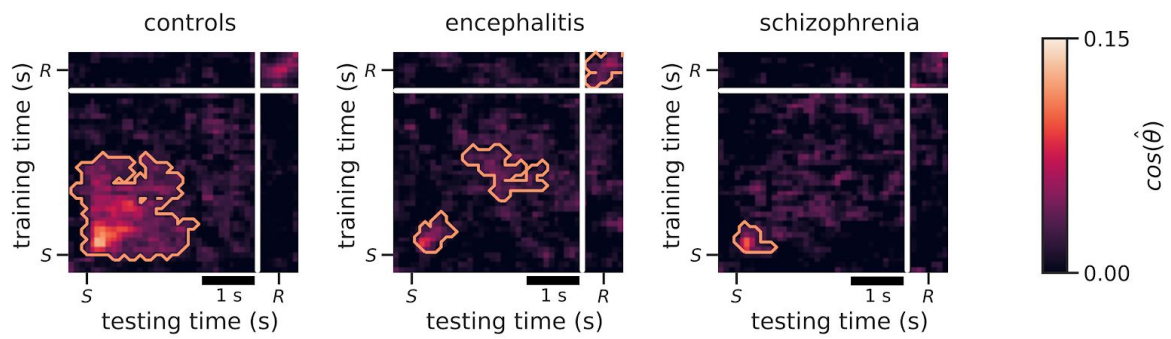
Supplementary Fig. 2 | Decoding performance as a function of the number of trials

To understand whether reduced memory codes in patients were related to the lower average number of trials in patients (number of trials: ctrl, $n = 900.63 \pm 103.02$; enc, $n = 870.44 \pm 141.93$; schz, $n = 833 \pm 153.33$), we measured the increase in decoding performance when including increasing numbers of trials. For each subject (transparent lines), we trained and tested decoders in mid-delay (0.65 - 0.85 s after stimulus onset) on the first n trials of a session (x-axis). **a-c** The upper row shows the increase in decoding performance in all participants, up to the minimum number of trials ($n=440$ trials) that any included subject had performed (ctrl, $n = 22$; enc, $n = 27$; schz, $n = 19$). **d-e** The lower row shows the increase in decoding performance in a subset of subjects (ctrl, $n = 19$; enc, $n = 21$; schz, $n = 10$) who had performed the full experiment (after excluding trials during preprocessing, the cutoff was set to $n = 880$ trials). Left (black), healthy controls, middle (green), encephalitis, right (purple), schizophrenia. In all groups, mean decoding performance across subjects increased when including more trials, but saturated at ~ 400 trials. Importantly, group differences remained marked when fixing the number of trials along the entire x-axis. Shading, 95% bootstrap C.I.



Supplementary Fig. 3 | Alpha power is not modulated less efficiently in patients than in healthy controls

We tested whether decreased decoding performance in patients could be an effect of a less efficient modulation of alpha power during working memory delays. To this end, we decomposed EEG signals at each electrode into different frequency bands using a multi-taper-method convolution (with fieldtrip function “mtmconvol”) with Hanning tapers. For each time point, we then calculated the $\log(\text{power}/f)$ and subtracted the average pre-stimulus signal (-0.5 - 0 s). **a-c**, All three groups show a significant modulation of spectral power across a wide range of frequencies (white contours, significant decrease; black contours, significant increase w.r.t. baseline, 1-sample cluster permutation test at $\alpha = 0.05$) after stimulus onset (0s - 1.25 s) and in the ITI (~ 1.5 - 0.5 s before stimulus onset). **d**, When comparing relative log power decreases between groups, there are no significant clusters of differences between ctrl and enc, or **e**, between ctrl and schz (2-sample cluster permutation test at $\alpha = 0.05$). **f**, When comparing the delay period (0.65 - 0.85 s) without correcting for multiple comparisons, there is a significant difference at 12 Hz between ctrl and enc. Group differences in relative log power during delay are in the opposite direction than group differences in decoding performance: Encephalitis patients modulate power in the alpha band more strongly than healthy controls (uncorrected permutation test at $\alpha = 0.05$). During delay, there is no difference in relative log power between ctrl and schz. **g**, In the fixation period (0.95 - 0.75 s before stimulus onset), low frequencies (2 - 4 Hz) are modulated more strongly in healthy controls than in both patient groups (uncorrected permutation test at $\alpha = 0.05$). Moreover, power at 14 Hz is modulated more strongly in encephalitis patients than in healthy controls. Grey bars indicate the alpha band, which was used in the decoding procedure.



Supplementary Fig. 4 | Cross-temporal decoding in 3 s delay trials

Cross-temporal decoding during stimulus presentation (S), delay, and response (R), trained and tested only on 3 s trials. Compare to Fig. 3d-f, where decoders were trained on pooled 1 and 3 s trials. White lines mark the discontinuity of EEG signals after the probe onset at 3.25 s. Orange lines mark significant decoding clusters (1-sample permutation test at $\alpha = 0.05$). Spatial locations were decoded above chance from $\sim 0 - 2$ s after stimulus onset in healthy controls. In patients, using only 3 s trials impaired decoding performance more dramatically, so that locations could only be decoded reliably around the time of stimulus presentation, and weakly around ~ 2 s for patients with encephalitis.

A4 Ten simple rules for modern psychophysics

In this section²⁴, I provide a practical guide to researchers who want to embark on their own psychophysics journey.

The section summarizes practical lessons learned during the work leading to this thesis. It is an overview with simple best practices for anyone who wants to set up a behavioral lab experiment in humans. These rules are useful not only for classical psychophysics, but to test hypotheses about how the brain performs perception, working memory, motor control, decision making, or any other abstract brain function in a task with a trial structure.

²⁴ This section is a manuscript in preparation. It is a joint project that started in discussions with Joao Barbosa and Alexandre Hyafil, and received additional inputs from Christopher Summerfield and Salvador Soto-Faraco.

Ten Simple Rules For Modern Psychophysics

Introduction

We are witnessing a major technological revolution in the field of neuroscience, with increasingly large-scale neurophysiological recordings in behaving animals (Gao and Ganguli 2015) and advanced techniques and computational methods for human neuroimaging, combined with the monitoring of high-dimensional behavior (Krakauer et al. 2017; Musall et al. 2019) and causal interventions (Jazayeri and Afraz 2017) at its forefront.

Despite being a relatively low-tech enterprise, modern psychophysics remains a fundamental way of investigating the mysteries underlying the human mind (Read 2015), especially when combined with computational modelling (Wilson and Collins 2019; Ma and Peters 2020), and is an affordable and accessible approach to neuroscience.

Here, rather than focussing on the underlying theory of psychophysics, we aim at providing a practical guide on how to perform successful experiments. These are the unwritten rules to a steady and successful walk through the workflow of a typical psychophysics experiment (Figure 1). For readers in search of a formal introduction to psychophysics, we refer to several seminal introductory handbooks (Lu and Doshier 2013; Kingdom and Prins 2016; Green and Swets 1989) or tutorial articles (Wichmann and Jäkel 2018; Wichmann and Hill 2001b).

We intentionally employ a wide definition of psychophysics, beyond the traditional, more narrow understanding which strictly refers to the study of the perceptual system alone (Gescheider 2013). Over the years, this definition has been gradually extended to include more cognitive tasks relating to higher level processes (Waskom et al. 2019). Here, we include in its definition any behavioral experiment where human subjects respond, through stereotyped behavior, to the controlled presentation of stimuli of any modality. With this broad definition in mind, we believe our rules can be useful for anyone starting a behavioral experiment to study working memory (Wilken and Ma 2004; Ma et al. 2014), long term memory (Batchelder and Riefer 1990), reinforcement learning (Sutton and Barto 2018), motor control (Wolpert et al. 2011; Gallivan et al. 2018), continuous psychophysics (Bonnen et al. 2015), and so on.

We assume here that the reader already has a hypothesis to test. Developing this hypothesis is the most creative part of psychophysics and we have no recipe for that. Instead, we want to provide clear, practical rules on how to proceed once you have identified a scientific question, have operationalised your hypothesis, and want to test it with psychophysics. The rules provided below can be seen as consecutive steps to take to get the behavioral dataset that will best test your hypothesis. During this process, some of the steps can be taken in parallel, while others are better taken iteratively in a loop, as shown in Figure 1. So how do you know you have a valid research question? Have someone ask you: "what is your question?". If you can't answer, then think again. If you can, have them ask you: "why is that important?". If you can't answer, go back to the

drawing board. By the way, "because nobody has done it before" is not a valid answer to this question.

Rule 1. Do it.

Psychophysics is one of the most affordable experimental disciplines. This, however, has not always been the case. Avant-garde psychophysical experiments dating back to the late 40s involved expensive technology that was difficult to fit in an office room (Koenderink 1999). Nowadays, running a typical psychophysics experiment requires only inexpensive equipment, a few hundred euros to compensate voluntary subjects, and a good hypothesis on how the brain processes information. Indeed, behavioral experiments on healthy human adults can be substantially faster and cheaper than other neuroscience experiments, such as human neuroimaging or experiments with other animals. In addition, ethical approval is easier to get (but see Rule 5), since psychophysics is the least invasive approach to study the mind, and subjects participate voluntarily. With some experience and a bit of luck, you could code your experiment, collect and analyze the data in a few weeks.

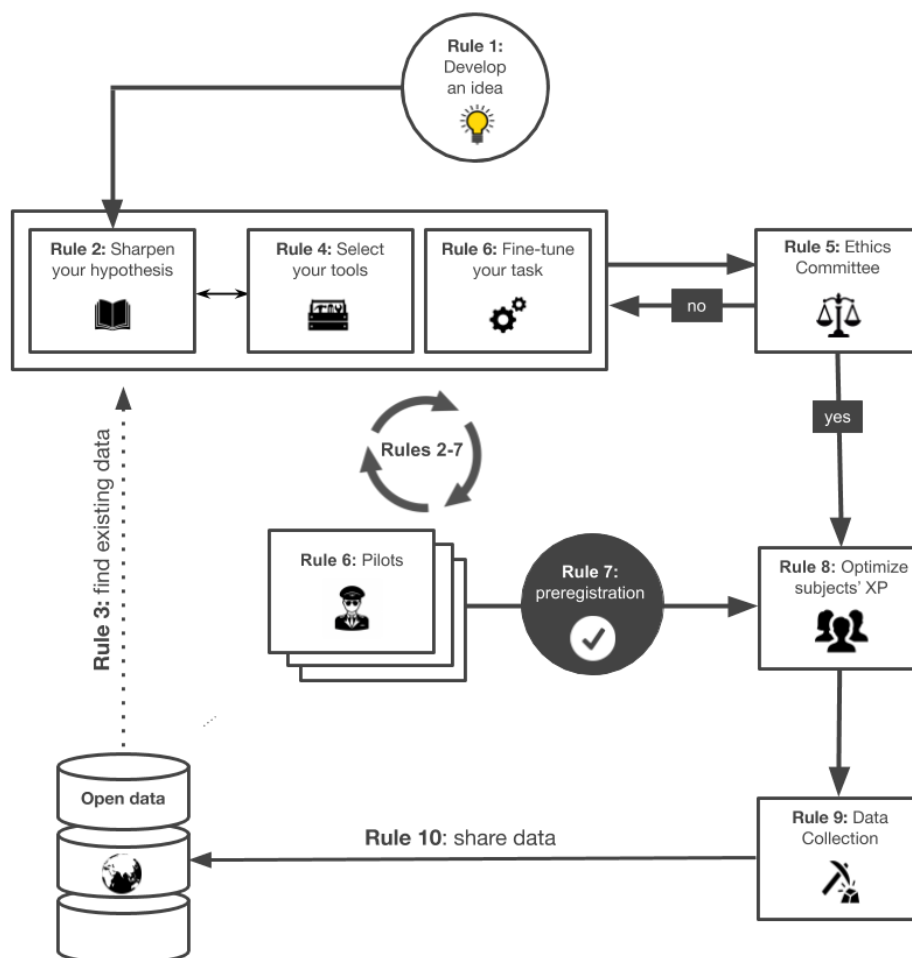


Figure 1. workflow of a successful psychophysics experiment.

However, you should not rush into data collection. A trivial question, an erroneous operationalization of the hypothesis, lack of statistical power, or a carelessly developed set of predictions may result in findings that are irrelevant to your original question, uninformative, or inconclusive. To achieve the necessary level of control, you will probably need to spend a couple of months polishing your experimental paradigm, especially if it is innovative. But rather than spending a long time exploring a potentially infinite set of task parameters, loop between Rules 2-7 until you converge on a solid design.

Rule 2. Aim for the best experimental design to test your hypothesis.

Once you have a good sense of the hypothesis and variables you want to measure in order to answer your question, it is time to become concrete on the type of task, the nature of stimuli to be used, which conditions to include, and whether they will be manipulated within or across subjects. Some of these aspects may be constrained by your question or hypothesis. For example, if you study speech the stimuli many times are bound to be acoustic. If gender is a variable of interest, you are forced to use a between-subject design. Think hard on which set of conditions will better address your hypothesis. As Albert Einstein famously did *not* say, “everything should be made as simple as possible, but not simpler”. That is a good mantra to keep in mind throughout the whole process, and especially during this stage. You should only manipulate a few variables of interest, which influence behavior in a way that is specific to the hypothesis under scrutiny. If your hypothesis unfolds into a series of sub-questions, focus on the core questions. A common mistake by newcomers is to design complex paradigms aimed at addressing too many questions, with dramatic repercussions on the statistical power. A rich set of experimental conditions, given sufficient number of trials, will provide better insight into cognitive processes if you master the necessary analysis tools to capture complex structures in your data. For example, if one wants to manipulate the difficulty of the task on a trial-by-trial basis, think of what is the most effective parameter to do so (e.g. if you run a perceptual task, think which parameter dial up or down the likely sources of noise). Don’t be afraid to innovate if you think this will provide better answers to your scientific questions. However, innovative paradigms can take much more time to adjust than using off-the-shelf solutions, so make sure the potential gain justifies the cost. Aim at experimental designs with orthogonal manipulations, to avoid confounds that will be difficult to control for *a posteriori* (Waskom et al. 2019; Dykstra 1966).

Another common mistake is to overlook the specific analyses that will be necessary to test your hypothesis before collecting the data, potentially leading to insufficient or inadequate data for the comparisons of interest (see Rule 10). Relatedly, simulate your hypothesized effect (Wilson and Collins 2019), with some assumptions about the noise in your measured behavior. This is useful to validate your analyses and will help decide the sample size of your experiments *a priori*, formalized in what is called a *power analysis* (Bausell and Li 2002). Alternatively, you might aim for sample sizes similar to previous studies, but bare in mind that you might be overestimating the effect size (Button et al. 2013; Kvarven et al. 2020).

Finally, it is easy to get over-excited about an idea, so ask your colleagues for feedback - even better, ask explicitly for *advice* (Yoon et al. 2019). You can do that through lab meetings or contacting your most critical collaborator that is good at generating alternative explanations - the more boring the alternative explanations, the more seriously you should take them. In sum, you should not cling to one idea, instead be your own critic and think of all the ways the experiment can fail - odds are it will. Before embarking on a project, look for any red flags that tell you to change substantially or maybe even abandon the project.

Table 1. Open repositories of behavioral data.

Database	Type of data	URL
Generic data		
DataverseNL	All types of data, including behavior	dataverse.nl
Dryad	Data from different fields of biology, including behavior	datadryad.org
Figshare	All types of data, including behavior	figshare.com
GIN	All types of data, including behavior	gin.g-node.org
Google Dataset Search	All types of data, including behavior	datasetsearch.research.google.com
Harvard Dataverse	All types of data, including behavior	dataverse.harvard.edu
Nature Scientific Data	All types of data, including behavior	nature.com/sdata
Neuroscience Information Framework	Meta-search tool for data, models and software from all areas of neuroscience	neuinfo.org
OpenLists	All types of electrophysiology, including behavior	github.com/openlists/ElectrophysiologyData
OSF	All types of data, including behavior and neuroimaging. Preregistration service	osf.io
Zenodo	All types of data, including behavior	zenodo.org
Human data		
APA	Shared data are available for use in psychological science research, curated by APA	apa.org/research/responsible/data-links
CamCan	Cognitive and neuroimaging data of subjects across adult lifespan	cam-can.org
Confidence database	Behavioral data with confidence measures	osf.io/s46pr
Human Brain Project	Mostly human and mouse recordings, including behavior	kg.ebrains.eu/search
Oasis	Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer's Disease	oasis-brains.org
Open Neuro	Human neuroimaging data, including behavior	openneuro.org
PsychArchives	All fields of psychology	psycharchives.org
The Healthy Brain Network	Psychiatric, behavioral, cognitive, and lifestyle phenotypes, as well as multimodal brain imaging of children and adolescents	fcon_1000.projects.nitrc.org/indi/cmi_health_y_brain_network
UCLA Library	Psychology data repositories compiled by trusted psychological authorities.	guides.library.ucla.edu/psychology/data
Animal data		
CRCNS	Animal behavior and electrophysiology	crcns.org

International Brain Lab	Mouse electrophysiology and behavior	data.internationalbrainlab.org
Mouse Bytes	Mouse cognition, imaging and genomics	mousebytes.ca

Rule 3. Check for existing datasets.

Before running new experiments, check for existing data that you could use, even if only to test simpler versions of your hypothesis. Researchers are increasingly open to sharing their data (see Rule 10), either publically or upon request. If the data from a published article is not publicly available (e.g. check the data sharing statement in the article), don't be afraid of writing an email to the corresponding author to politely ask for data, that is what the contact email is for. In the best case scenario, you could find the perfect data to address your hypothesis, without having to collect it - but beware that data decay can become a problem (Thompson et al. 2019). If you cannot find exactly what you were looking for, playing with data from similar experiments will still help you get a feeling for the kind of data you can obtain, potentially suggesting ways to improve your own experiment. In Table 1 you can find several repositories with all sorts of behavioral data, both from human subjects and other species, often accompanied by neural recordings.

Rule 4. Choose the right equipment and environment.

If the hypothesis you are testing needs you to control for things such as luminosity, noise, eye-movement, precise timing of events or the exact placement of hardware, then you should run your experiment in an appropriate lab. Typical psychophysical set-ups consist of a small room in which you can ideally control, or at least measure, these factors. If your experiment is long, it is particularly important that the subject sits comfortably (see Rule 8). In the lab, you can ensure that experiments will not be interrupted (e.g. by the subject's mobile phone, lab mates or other people working closeby). Especially, think of all possible confounds for your variable of interest. For example, if eye movements can be a confound, you can attempt to control by design, by eliminating the incentive of moving the eyes or, if not possible, use an eye tracker. There are several affordable options, including some that you have to build from scratch (Hosp et al. 2020; Mantiuk et al. 2012), and work reasonably well if ensuring fixation is all you need (Funke et al. 2016). If your lab has an EEG setup, you can record EOG signals to have rough measures of eye movements (e.g. Quax et al. 2019). For open source, low-cost "brain and body sensors", check OpenBCI products (Frey 2016). Before buying any expensive equipment, check if someone in your community already has the tool you need, and importantly, if it is compatible with the rest of your toolkit, such as response devices, available ports, eye trackers, but also software, and your operating system.

The "keep it simple" mantra (Rule 2) is important also during this stage. If you conclude that some or all of the above factors won't affect your results, consider running your experiment on more flexible and maybe cheaper devices, such as tablets (Linares et al. 2018). Alternatively, you can take it online. An online experiment allows you to scale things up (Stewart et al. 2017; Difallah et al. 2018), effectively speeding up data collection by orders of magnitude - but it comes at the cost of losing experimental control (Crump et al. 2013),

and data collected online is much noisier than when collected in the lab. Additionally, the average low payment in crowdsourcing platforms biases us to pay less than what is ethically correct. For example, data-hungry companies are known to exploit vulnerable populations in online platforms (Alana Semuels 2018). A minimum should be defined by the minimum wage in the country you recruit your subjects from. It will likely be above the average payment in the platform, but still cheaper than running the experiment in the lab, where you would have to pay both the subject and the experimenter. Paying well is not only ethically correct, but it will allow you to filter for best performers and ensure faster and higher data quality (Stewart et al. 2017).

Finally, choose a good programming language, especially if you do not have strong preferences - yet. Python, for example, is open source, free, versatile and becoming the go-to language in data science (Kaggle 2019) with plenty of tutorials for all levels of proficiency. PsychoPy is a great option to implement your experiment, should you choose to do it in Python. If you are a proficient Matlab user, Psychtoolbox (Borgo et al. 2012) is a great tool, too. If you are considering running your experiment on a tablet or even a smartphone, you could use StimuliApp (Marin-Campos et al. 2020). Otherwise check Hans Strasburger's page (Strasburger 1994) that has provided a comprehensive and up-to-date overview of different tools, among other technical tips, for the last 25 years.

Rule 5. Submit early to the ethics committee.

This is a mandatory, yet tedious step. Do it early to avoid surprises that could halt your progress. Depending on the institution, the whole process can take several months, which you can use for piloting (see Rule 6). In your application, describe your experiment in terms general enough as to accommodate for changes in the paradigm that will inevitably occur. This is of course without neglecting potentially relevant ethical issues, especially if your target population includes vulnerable populations (such as patients or minors of age). You will have to describe factors concerning the sample, such as the details of participant recruitment planned and justified sample sizes (see Rule 3), and how the subjects' data will be anonymized and protected. You should also provide insight into the details of the experiment, including its duration, the experimental setup or platform (some ethics committee might not allow crowdsourcing experiments, for example), potential physiological recordings or interventions and whether they could harm the subjects in any way or how the subjects will be compensated for their participation (e.g. whether performance-dependent rewards will be used). You should also provide the consent form that you will ask participants to sign. Each committee has specific requirements, so ask colleagues for their documents and experiences, and start from there - often, the basic elements of an ethics application are widely recyclable, and this is the one case in research where copy-pasting is highly recommendable. Finally, keep in mind that as you want to go through the least rounds of review as possible, you should make sure you are abiding by all the rules.

Rule 6. Polish your experimental design through piloting.

Take time to run several pilots while fine-tuning your task parameters, especially the changes with respect to previous studies. Pilot yourself first. Then, pilot your friends and colleagues - chances are they will do the same to you, if they didn't already. In some cases, it is worth considering using online platforms to run pilot studies, especially when you want to sweep through many parameters (but see Rule 5). Typical parameters that need piloting to adjust are the size and duration of stimuli, masking, the duration of the inter-trial-interval, the modality of the response and feedback, and so on. Using values from previous studies can spare you some piloting. For example, if you plan to use a fixation cross, use the best one (Thaler et al. 2013). You should perform *quality checks* on pilot data, making sure that typical findings are confirmed (e.g. higher accuracy and faster reaction times for easier trials, post-error slowing of responses, preference of higher rewards in economic decision making, etc.), that most responses occur within the allowed time window, etc. These sanity checks can form the basis for what your exclusion criteria will be, e.g. applying cutoff thresholds on the proportion of correct trials, proportion of very fast responses (e.g. below 200 ms), lapse rate, etc. Make sure your exclusion criteria are always orthogonal to your main question. You can decide the criteria after you collect a cohort of subjects but always make decisions about whom (or which trials) to exclude *before* testing the main hypotheses in that cohort. By doing these sanity checks, you will also reveal potential bugs in your code, such as incorrectly saved data or an incorrect assignment of stimuli to task conditions (Table 2).

Find the right pace for the experiment to avoid boredom, tiredness or impulsive responses. Make conscious decisions on what aspects of the task can be fixed-paced or self-paced. At this stage you should also pay attention to the *lapse rate* of your pilots, a proxy of task engagement (Fetsch 2016) (see also Rule 8). A high lapse rate might, depending on the goals of your experiment, reflect random responses from that subject (Fetsch 2016), which would introduce noise in your dataset and generally does not involve a cognitive process of interest (but see Pisupati et al. 2019). If you are interested in computing psychometric curves, you should sample stimuli from a broad range of difficulties (Rule 2) (Waskom et al. 2019), including the easiest ones if you want to compute lapse rates, for example to compare different populations (Linares et al. 2019). If you want to make sure all subjects perform at similar levels of performance, or if you are interested in studying individual psychophysical thresholds, consider using a staircase procedure (Cornsweet 1962; Kingdom and Prins 2016) to adjust task difficulty for each subject, or alternatively a psi-marginal adaptive method (Prins 2013), that takes attention lapses into account. Relatedly, make sure that the performance of your pilots is reasonably stable across experimental sessions or blocks, so that learning or fatigue does not play a substantial role. You should also decide whether to sample the stimulus sequence randomly or in a balanced way (i.e., sampling with or without replacement from the pool of possible stimuli). In general, balanced sampling is the best option, since unbalanced stimulus sequences can introduce confounds difficult to control *a posteriori* (Dykstra 1966). When imposing a balanced sequence, make sure stimulus predictability will not confound your interpretation.

If your design really needs truly random sequences, make sure you have enough trials to avoid strong imbalances due to undersampling.

Finally, make sure you will be using your resources wisely. For example, should you run few subjects with many trials or many subjects with relatively few trials? There is no silver bullet for these questions, as it depends largely on the type of question you are trying to answer. A rule of thumb is that if you are interested in studying different strategies or other individual characteristics (e.g. Tversky and Kahneman 1974), then you should focus on collecting data from as many subjects as possible (Waskom et al. 2019), but beware that even large numbers of subjects can lead to poor test-retest reliability (Elliott et al. 2020). On the other hand, classical psychophysics is interested in the computational principles behind a given perceptual or cognitive process that are common to all normally-functioning humans (Read 2015) and therefore requires subjects whose performance is stable, and whose performance threshold is thoroughly sampled with many trials (Smith and Little 2018; Waskom et al. 2019).

Table 2. Top 10 most common coding and data handling errors committed by the authors when doing psychophysics and how to avoid them. These are loosely sorted by type of error (crashes, incorrect runs, saving data issues), not by frequency.

Common mistake	How to avoid
1) Code breaks in the middle of a session, and all data is lost.	Save your data at the end of each block or, if possible, at the end of each trial.
2) Your code breaks when a certain key is hit, or when secondary external hardware (e.g. eye tracker) unexpectedly stops sending signals.	Check which keys are assigned in your paradigm, and which lead to the interruption of the program. Check in advance what happens if external hardware problems emerge. Perform a <i>crash test</i> of your code to make sure it is resilient to wrong keys being hit, or keys being hit at the wrong time.
3) You made an “improvement” just before the experimental session. Your code now breaks unexpectedly or doesn’t run at all during data collection.	Avoid using untested code.
4) Some software sends notifications, such as software updates, in the middle of a session. The experiment is interrupted, and the subject might not even notify you.	Switch off all software you don’t need, disable automatic updates. Disable the internet connection.
5) The randomization of stimuli or conditions is wrong, or identical for all subjects.	Make sure to use a different seed whenever you want your data to be independent. After piloting, inspect the distribution of conditions in your data.
6) Your subject is not doing what they should and you don’t notice.	Have a control screen or a remote connection to mirror the subject’s display (e.g. with Chrome Remote Desktop), but make sure it will not introduce delays. There, also print ongoing performance measures.
7) You overwrite data/code from earlier sessions or subjects. This data is now lost.	Add a line of code that checks if a filename already exists before writing. Backup output directory regularly through git. Alternatively or additionally, save your data directly on dropbox, google drive or another automatic system.
8) You save participant data with the wrong identifier and later cannot assign it correctly.	Use multiple identifiers to name a file: subject and session ID + date and time + computer ID, for example.

9) You decided at some point to adjust “constant” experimental parameters during data collection. Now, which participants saw what?	Define all experimental parameters at the beginning of your code, preferably in a flexible format such as a python <i>dictionary</i> , and save them in a separate log file for each session or include them in your table repeatedly for each trial.
10) After data collection, you start having doubts about the timing of events, and the temporal alignment with continuous data, possibly stored on another device (fMRI, eye tracking).	Save “time stamps” in your data table for each event of a trial (fixation onset, stimulus onset, etc.). Make sure your first event is temporally aligned to the onset of continuous data.

Rule 7. Preregister or replicate your experiment.

An alarming proportion of researchers in the field of psychology reports to have been involved in some form of questionable research practices (John et al. 2012; Fiedler and Schwarz 2016). The line between right and wrong here is, at best, a blurry one. Sometimes, making the right or the wrong call needs a degree of statistical sophistication, and even so, we are humans and our decisions are often not solely determined by rationality. Two common forms of questionable practices, p-hacking and HARKing (Stroebe et al. 2012), increase the likelihood of obtaining false positive results. In p-hacking (Simmons et al. 2011), significant tests are not corrected for the multiple alternative hypotheses tests that were performed, while harking refers to the formulation of a hypothesis after the results are known (Kerr 1998). It is sometimes very difficult to avoid the temptation to decide that, in fact, what we need to use as the dependent variable is the median, after we have seen that the mean gave an unsatisfactory outcome. Additionally, the journals where we aim at publishing have a traditional bias for positive findings, while our negative results often remain unpublished (Rosenthal 1979). These practices posit a substantial threat to the efficiency of research, and they are believed to underlie the replication crisis in psychology (Open Science Collaboration 2015). Ironically, the failure to replicate is highly replicable (Klein et al. 2018; Klein et al. 2014).

This crisis has motivated the practice of preregistering experiments before the actual research is conducted (Kupferschmidt 2018). In practice, this consists of a short document that answers standardized questions about the experimental design and planned statistical analyses. The optimal time for preregistration is once you finish tweaking your experiment through piloting and power analyses (Rule 2-6). Preregistration may look like an extra hassle before finally collecting data, but it will actually often save you time: by forcing you to write down explicitly all your analyses, you may find some inconsistencies and go back to your paradigm. Additionally, the text you generate at this point can be reused for the methods section of your manuscript. Alternatively, you can opt for registered reports, where you submit a prototypic version of your final manuscript, without the results, to peer-review (D. Stephen Lindsay 2016). If your report survives peer-review, it is accepted in principle, which means that whatever the outcome, the manuscript will be published (given that the study was rigorously conducted). High-impact journals such as eLife, Nature Human Behavior, and Nature Communications already accept this format.

There are several databases that manage preregistrations, such as Open Science Framework (OSF.io), which is the most popular one, or AsPredicted.org, that offers more

concrete guidelines. Importantly, these platforms keep your registration private, so there is no added risk of being scooped. Preregistering your analyses does not mean you can't do exploratory analyses, just that these analyses will be explicitly marked as such. This transparency strengthens your arguments when reviewers read your manuscript and protects you from committing scientific misconduct involuntarily.

Rule 8. Take care of your subjects.

Remember that your subjects are volunteers, not your employees. In a way, they are helping science progress, so acknowledge that and treat them with respect (Wichmann and Jäkel 2018). Be generous with the compensation, it is also in your interest to avoid low turn-out. Send emails with enough information well in advance. Set up an online booking system, for example through Doodle, where subjects can select their preferred schedule from the available slots. This will avoid an unnecessary back and forth of emails. If you cannot rely on an existing database of participants for recruitment, create one and ask your participants for permission to include them in it (make sure to comply with the regulations on data protection). Try to maintain a long and diverse list, but eliminate unreliable participants.

Make sure your subjects come fully awake, healthy and without the influence of any drugs or medication that might alter their perception or performance. Systematize a routine to when participants arrive at the lab. It is very easy for participants to confuse the meaning of events within a trial (e.g. what is fixation, cue, stimulus, response prompt, feedback), especially if they occur in rapid succession. To avoid this, write a clear and concise instruction sheet and have your participants read it before the experiment. This will make sure they have all the needed information and will avoid framing effects (Tversky and Kahneman 1989). Allow time for clarifying questions and repeat instructions on the screen during the corresponding stages of the experiment (introduction, practice block, break, etc). If a collaborator is collecting the data for you, spend some time training them and designing a clear protocol (e.g. checklist including how to calibrate the eye tracker), including troubleshooting. Be their first mock subject, and be there for their first real subject.

To ease subjects into the task, have a practice block with very easy trials that become progressively more similar to actual trials, for example by changing the event timings or the stimulus contrast. If you avoid overwhelm their attentional capacities, the subjects will more rapidly automatize parts of the process (e.g. cue-rule associations, button associations, etc.). Moreover, bad performance on these very easy trials could reflect the subjects' misunderstanding even the basic rules. Unless you have a reason to do otherwise, do short blocks allowing for quick breaks (humans get bored quickly) every ~5 minutes (Wichmann and Jäkel 2018). Include the possibility for one or two longer breaks, because some subjects might need more or longer breaks than others. Incentivize them to perform well if possible and approved by the ethics committee, for example by offering a bonus if they reach a certain performance level, but let them know it is normal to make errors. Even better, try to gamify your paradigm, for example by providing quick feedback after each trial, or block-wise scores. In general, we recommend giving performance feedback after

each trial. There are however circumstances where you want to avoid it, in particular when you think it might interfere with the cognitive process under study. We think this could happen in two situations. First, feedback is known to influence the next few trials due to win-stay-lose-switch strategies (Abrahamyan et al. 2016; Urai et al. 2017; Hermoso-Mendizabal et al. 2020) or other types of superstitious behavior (Ono 1987). This nuisance should have very limited impact on your analyses of interest, unless you are precisely interested in sequential effects (Hermoso-Mendizabal et al. 2020; Lak et al. 2020). Second, feedback can be used as a learning signal by the participants (Massaro 1969), but not much for low sensory detection tasks as it is for strategy-based paradigms or paradigms that include confidence reports (Schustek et al. 2019). Finally, at the end of every block of trials, display the remaining number of blocks - show them the *stairway to heaven*.

Rule 9. Record everything.

Once you have laid out what data you need to test your hypothesis, record everything else you can record. The dataset you are generating might be useful for others or your future self in ways you cannot predict now. Use an eye tracker if you need to ensure fixation, or if you are interested in rapid pupil dilatations as a proxy for information processing (Cheadle et al. 2014) or decision confidence (Urai et al. 2017). If you are using a mouse to register the subjects report, record all the mouse movements, reaction times, etc. Save your data in a tidy table format (Wickham 2014) (e.g. manipulate in pandas, but store as csv). Save your data as a table with one trial per line and all the relevant variables as columns (both presentation and behavioral variables), they are easier to analyze and to share. However, if some modality gives continuous output, such as pupil dilation or cursor position, save it in a separate file rather than creating *kafkaesque* data structures. If you use an eye-tracker or neuroimaging device, make sure you save synchronized timestamps in both data streams for later data alignment (see Table 2). Don't be afraid of having redundant columns (e.g. response and response accuracy), redundancy enables robustness to your mistakes. If you end up changing your paradigm, even if with small changes, save those version names in a lab notebook. If the lab does not use a lab notebook, start using one (Schnell 2015). Mark *all* incidents there, even those that seem uninteresting now. Back up your code and data regularly (see also Rule 10). Finally, don't stop data collection after the experiment is done: sometimes it is useful to include an informal questionnaire at the end, e.g. demographics (should you have approval from the ethics committee), but also important questions like "did you see so-and-so?" or "tell us about the strategy you used to solve part II".

Rule 10. Be transparent, and share your data.

Upon publication, share everything needed to replicate your findings in a repository or shared database (see Table 1). That includes your data and code. You should aim at properly documented code, but don't let that be the reason not to share. After all, bad code is better than no code (Barnes 2010; Gleeson et al. 2017). If possible, avoid proprietary software. Use python or R notebooks (Rule et al. 2019) to develop your analyses and git for version control (Perez-Riverol et al. 2016). Notebooks make code sharing easier with the community, but also with advisors or colleagues, when asking for help.

References

- Abrahamyan, A., Silva, L.L., Dakin, S.C., Carandini, M., Gardner, J.L. 2016. Adaptable history biases in human perceptual decisions. *Proceedings of the National Academy of Sciences of the United States of America*, 113(25), pp. 3548-3557.
- Alana Semuels 2018. The Online Hell of Amazon's Mechanical Turk [Online]. Available at: <https://www.theatlantic.com/business/archive/2018/01/amazon-mechanical-turk/551192/> [Accessed: 30 April 2020].
- Barnes, N. 2010. Publish your computer code: it is good enough. *Nature* 467(7317), p. 753.
- Batchelder, W.H. and Riefer, D.M. 1990. Multinomial processing models of source monitoring. *Psychological Review* 97(4), pp. 548–564.
- Bausell, R.B. and Li, Y.-F. 2002. *Power analysis for experimental research: A practical guide for the biological, medical and social sciences*. Cambridge: Cambridge University Press.
- Bonnen, K., Burge, J., Yates, J., Pillow, J. and Cormack, L.K. 2015. Continuous psychophysics: Target-tracking to measure visual sensitivity. *Journal of Vision* 15(3).
- Borgo, M., Soranzo, A. and Grassi, M. 2012. Psychtoolbox: sound, keyboard and mouse. In: *MATLAB for Psychologists*. New York, NY: Springer New York, pp. 249–273.
- Button, K.S., Ioannidis, J.P.A., Mokrysz, C., et al. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience* 14(5), pp. 365–376.
- Campbell, F.W. and Robson, J.G. 1968. Application of Fourier analysis to the visibility of gratings. *The Journal of Physiology* 197(3), pp. 551–566.
- Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., de Gardelle, V., Hecce Castañón, S. and Summerfield, C. 2014. Adaptive gain control during human perceptual choice. *Neuron*, 81(6), pp. 1429–1441.
- Cornsweet, T.N. 1962. The Staircase-Method in Psychophysics. *The American journal of psychology* 75(3), p. 485.
- Crump, M.J.C., McDonnell, J.V. and Gureckis, T.M. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *Plos One* 8(3), p. e57410.
- D. Stephen Lindsay, D.J.S., Scott O. Lilienfeld 2016. Research Preregistration 101 – Association for Psychological Science – APS. *APS Observer*.
- Difallah, D., Filatova, E. and Ipeirotis, P. 2018. Demographics and dynamics of mechanical turk workers. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18*. New York, New York, USA: ACM Press, pp. 135–143.
- Dykstra, O. 1966. The orthogonalization of undesigned experiments. *Technometrics: a journal of statistics for the physical, chemical, and engineering sciences* 8(2), p. 279.
- Elliott, M.L., Knodt, A.R., Ireland, D. et al. 2020. What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, 31(7), pp. 792-806.
- Fetsch, C.R. 2016. The importance of task design and behavioral control for understanding the neural basis of cognitive functions. *Current Opinion in Neurobiology* 37, pp. 16–22.
- Fiedler, K. and Schwarz, N. 2016. Questionable Research Practices Revisited. *Social psychological and personality science* 7(1), pp. 45–52.
- Frey, J. 2016. Comparison of an Open-hardware Electroencephalography Amplifier with Medical Grade Device in Brain-computer Interface Applications. In: *Proceedings of the 3rd International Conference on Physiological Computing Systems*. SCITEPRESS - Science and Technology

Publications, pp. 105–114.

Funke, G., Greenlee, E., Carter, M., Dukes, A., Brown, R. and Menke, L. 2016. Which eye tracker is right for your research? performance evaluation of several cost variant eye trackers. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60(1), pp. 1240–1244.

Gallivan, J.P., Chapman, C.S., Wolpert, D.M. and Flanagan, J.R. 2018. Decision-making in sensorimotor control. *Nature Reviews. Neuroscience* 19(9), pp. 519–534.

Gao, P. and Ganguli, S. 2015. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current Opinion in Neurobiology* 32, pp. 148–155.

Gescheider 2013. *Psychophysics: The Fundamentals*. Psychology Press.

Gleeson, P., Davison, A.P., Silver, R.A. and Ascoli, G.A. 2017. A commitment to open source in neuroscience. *Neuron* 96(5), pp. 964–965.

Green, D.M. and Swets, J.A. 1989. *Signal Detection Theory and Psychophysics*. Peninsula Pub.

Hermoso-Mendizabal, A., Hyafil, A., Rueda-Orozco, P.E., Jaramillo, S., Robbe, D. and de la Rocha, J. 2020. Response outcomes gate the impact of expectations on perceptual decisions. *Nature Communications* 11(1), p. 1057.

Hosp, B., Eivazi, S., Maurer, M., Fuhl, W., Geisler, D. and Kasneci, E. 2020. RemoteEye: An open-source high-speed remote eye tracker : Implementation insights of a pupil- and glint-detection algorithm for high-speed remote eye tracking. *Behavior research methods*.

Jazayeri, M. and Afraz, A. 2017. Navigating the neural space in search of the neural code. *Neuron* 93(5), pp. 1003–1014.

John, L.K., Loewenstein, G. and Prelec, D. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23(5), pp. 524–532.

Kaggle 2019. State of Data Science and Machine Learning 2019 [Online]. Available at: <https://www.kaggle.com/kaggle-survey-2019> [Accessed: 16 May 2020].

Kerr, N.L. 1998. HARKing: hypothesizing after the results are known. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc* 2(3), pp. 196–217.

Kingdom, F. and Prins, N. 2016. *Psychophysics*. Elsevier.

Klein, R.A., Ratliff, K.A., Vianello, M., et al. 2014. Investigating Variation in Replicability. *Social psychology* 45(3), pp. 142–152.

Klein, R.A., Vianello, M., Hasselman, F., et al. 2018. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in methods and practices in psychological science* 1(4), pp. 443–490.

Koenderink, J.J. 1999. Virtual Psychophysics. *Perception* 28(6), pp. 669–674.

Krakauer, J.W., Ghazanfar, A.A., Gomez-Marin, A., MacIver, M.A. and Poeppel, D. 2017. Neuroscience needs behavior: correcting a reductionist bias. *Neuron* 93(3), pp. 480–490.

Kupferschmidt, K. 2018. More and more scientists are preregistering their studies. Should you? *Science*.

Kvarven, A., Strømland, E. and Johannesson, M. 2020. Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour* 4(4), pp. 423–434.

Lak, A., Hueske, E., Hirokawa, J., et al. 2020. Reinforcement biases subsequent perceptual decisions when confidence is low, a widespread behavioral phenomenon. *eLife*, 9, e49834

Linares, D., Amoretti, S., Marin-Campos, R., et al. 2019. Perceptual spatial suppression and sensitivity for motion are weakened in schizophrenia. *BioRxiv*.

- Linares, D., Marin-Campos, R., Dalmau, J. and Compte, A. 2018. Validation of motion perception of briefly displayed images using a tablet. *Scientific Reports* 8(1), p. 16056.
- Lu, Z. and Doshier, B. 2013. *Visual Psychophysics: From Laboratory To Theory*. Cambridge, Massachusetts: The Mit Press.
- Mantiuk, R., Kowalik, M., Nowosielski, A. and Bazyluk, B. 2012. Do-It-Yourself Eye Tracker: Low-Cost Pupil-Based Eye Tracker for Computer Graphics Applications. *Lecture Notes in Computer Science (Proc. of MMM 2012)* 7131, pp. 115–125.
- Marin-Campos, R., Dalmau, J., Compte, A., Linares, D. 2020. StimuliApp: psychophysical tests on mobile devices. *PsyArXiv*.
- Massaro, D.W. 1969. The effects of feedback in psychophysical tasks. *Perception & Psychophysics* 6(2), pp. 89–91.
- Ma, W.J., Husain, M. and Bays, P.M. 2014. Changing concepts of working memory. *Nature Neuroscience* 17(3), pp. 347–356.
- Ma, W.J. and Peters, B. 2020. A neural network walks into a lab: towards using deep nets as models for human behavior. *arXiv*.
- Musall, S., Urai, A.E., Sussillo, D. and Churchland, A.K. 2019. Harnessing behavioral diversity to understand neural computations for cognition. *Current Opinion in Neurobiology* 58, pp. 229–238.
- Ono, K. 1987. Superstitious behavior in humans. *Journal of the experimental analysis of behavior* 47(3), pp. 261–271.
- Open Science Collaboration 2015. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* 349(6251), p. aac4716.
- Perez-Riverol, Y., Gatto, L., Wang, R., et al. 2016. Ten simple rules for taking advantage of git and github. *PLoS Computational Biology* 12(7), p. e1004947.
- Pisupati, S., Chartarifsky-Lynn, L., Khanal, A. and Churchland, A.K. 2019. Lapses in perceptual judgments reflect exploration. *BioRxiv*.
- Prins, N. 2013. The psi-marginal adaptive method: How to give nuisance parameters the attention they deserve (no more, no less). *Journal of Vision* 13(7), p. 3.
- Quax, S.C., Dijkstra, N., van Staveren, M.J., Bosch, S.E. and van Gerven, M.A.J. 2019. Eye movements explain decodability during perception and cued attention in MEG. *Neuroimage* 195, pp. 444–453.
- Read, J.C.A. 2015. The place of human psychophysics in modern neuroscience. *Neuroscience* 296, pp. 116–129.
- Rosenthal, R. 1979. The file drawer problem and tolerance for null results. *Psychological Bulletin* 86(3), pp. 638–641.
- Rule, A., Birmingham, A., Zuniga, C., et al. 2019. Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. *PLoS Computational Biology* 15(7), p. e1007007.
- Schnell, S. 2015. Ten simple rules for a computational biologist's laboratory notebook. *PLoS Computational Biology* 11(9), p. E1004385.
- Schustek, P., Hyafil, A. and Moreno-Bote, R. 2019. Human confidence judgments reflect reliability-based hierarchical integration of contextual information. *Nature Communications*, 10, e5430.
- Simmons, J.P., Nelson, L.D. and Simonsohn, U. 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11), pp. 1359–1366.
- Smith, P.L. and Little, D.R. 2018. Small is beautiful: In defense of the small-N design. *Psychonomic*

Bulletin & Review 25(6), pp. 2083–2101.

Stewart, N., Chandler, J. and Paolacci, G. 2017. Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences* 21(10), pp. 736–748.

Strasburger, H. 1994. Strasburger's psychophysics software overview [Online]. Available at: <http://www.visionscience.com/documents/strasburger/strasburger.html> [Accessed: 16 May 2020].

Stroebe, W., Postmes, T. and Spears, R. 2012. Scientific Misconduct and the Myth of Self-Correction in Science. *Perspectives on psychological science : a journal of the Association for Psychological Science* 7(6), pp. 670–688.

Sutton, R.S. and Barto, A.G. 2018. *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge, Massachusetts: A Bradford Book.

Thaler, L., Schütz, A.C., Goodale, M.A. and Gegenfurtner, K.R. 2013. What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research* 76, pp. 31–42.

Thompson, W.H., Wright, J., Bissett, P.G. and Poldrack, R.A. 2019. Dataset Decay: the problem of sequential analyses on open datasets. *BioRxiv*.

Tversky, A. and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185(4157), pp. 1124–1131.

Tversky, A. and Kahneman, D. 1989. Rational choice and the framing of decisions. In: Karpak, B. and Zionts, S. eds. *Multiple criteria decision making and risk analysis using microcomputers*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 81–126.

Urai, A.E., Braun, A., and Donner, T.H. 2017. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, 8, 14637.

Waskom, M.L., Okazawa, G. and Kiani, R. 2019. Designing and interpreting psychophysical investigations of cognition. *Neuron* 104(1), pp. 100–112.

Wichmann, F.A. and Hill, N.J. 2001a. The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics* 63(8), pp. 1293–1313.

Wichmann, F.A. and Hill, N.J. 2001b. The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics* 63(8), pp. 1314–1329.

Wichmann, F.A. and Jäkel, F. 2018. Methods in Psychophysics. In: Wixted, J. T. ed. *Stevens' handbook of experimental psychology and cognitive neuroscience*. Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 1–42.

Wickham, H. 2014. Tidy Data. *Journal of statistical software* 59(10).

Wilken, P. and Ma, W.J. 2004. A detection theory account of change detection. *Journal of Vision* 4(12), pp. 1120–1135.

Wilson, R.C. and Collins, A.G. 2019. Ten simple rules for the computational modeling of behavioral data. *eLife* 8.

Wolpert, D.M., Diedrichsen, J. and Flanagan, J.R. 2011. Principles of sensorimotor learning. *Nature Reviews. Neuroscience* 12(12), pp. 739–751.

Yoon, J., Blunden, H., Kristal, A.S. and Whillans, A.V. 2019. Framing Feedback Giving as Advice Giving Yields More Critical and Actionable Input. *Harvard Business School Working Paper*, No. 20-021, August 2019

Bibliography

Adam, K.C.S., Vogel, E.K. and Awh, E. 2017. Clear evidence for item limits in visual working memory. *Cognitive Psychology* 97, pp. 79–97.

Akrami, A., Kopec, C.D., Diamond, M.E. and Brody, C.D. 2018. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature* 554(7692), pp. 368–372.

Almeida, R., Barbosa, J. and Compte, A. 2015. Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. *Journal of Neurophysiology* 114(3), pp. 1806–1818.

Amit, D.J. and Brunel, N. 1997. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex* 7(3), pp. 237–252.

Anticevic, A., Gancsos, M., Murray, J.D., et al. 2012. NMDA receptor function in large-scale anticorrelated neural systems with implications for cognition and schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America* 109(41), pp. 16720–16725.

Arguello, P.A. and Gogos, J.A. 2012. Genetic and cognitive windows into circuit mechanisms of psychiatric disease. *Trends in Neurosciences* 35(1), pp. 3–13.

Arnsten, A.F.T., Wang, M.J. and Paspalas, C.D. 2012. Neuromodulation of thought: flexibilities and vulnerabilities in prefrontal cortical network synapses. *Neuron* 76(1), pp. 223–239.

Ashourian, P. and Loewenstein, Y. 2011. Bayesian inference underlies the contraction bias in delayed comparison tasks. *Plos One* 6(5), p. e19551.

Atkinson, R.C. and Shiffrin, R.M. 1968. Human Memory: A Proposed System and its Control Processes. In: *Psychology of Learning and Motivation*. Elsevier, pp. 89–195.

Attwell, D. and Laughlin, S.B. 2001. An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism* 21(10), pp. 1133–1145.

Avery, M.C. and Krichmar, J.L. 2017. Neuromodulatory systems and their interactions: A review of models, theories, and experiments. *Frontiers in Neural Circuits* 11, p. 108.

Badcock, J.C., Badcock, D.R., Read, C. and Jablensky, A. 2008. Examining encoding imprecision in spatial working memory in schizophrenia. *Schizophrenia Research* 100(1–3), pp. 144–152.

- Baddeley, A. 1992. Working memory. *Science* 255(5044), pp. 556–559.
- Baddeley, A. 1986. *Working Memory*. Oxford: Oxford University Press.
- Bae, G.-Y., Olkkonen, M., Allred, S.R. and Flombaum, J.I. 2015. Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General* 144(4), pp. 744–763.
- Banerjee, S., Snyder, A.C., Molholm, S. and Foxe, J.J. 2011. Oscillatory alpha-band mechanisms and the deployment of spatial attention to anticipated auditory and visual target locations: supramodal or sensory-specific control mechanisms? *The Journal of Neuroscience* 31(27), pp. 9923–9932.
- Barak, O. and Tsodyks, M. 2007. Persistent activity in neural networks with dynamic synapses. *PLoS Computational Biology* 3(2), p. e35.
- Barbieri, F. and Brunel, N. 2008. Can attractor network models account for the statistics of firing during persistent activity in prefrontal cortex? *Frontiers in Neuroscience* 2(1), pp. 114–122.
- Barbosa, J. and Compte, A. 2020. Build-up of serial dependence in color working memory. *Scientific Reports* 10(1), p. 10959.
- Barch, D.M. and Ceaser, A. 2012. Cognition in schizophrenia: core psychological and neural mechanisms. *Trends in Cognitive Sciences* 16(1), pp. 27–34.
- Bays, P.M., Catalao, R.F.G. and Husain, M. 2009. The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision* 9(10), pp. 7.1-11.
- Bliss, D.P. and D’Esposito, M. 2017. Synaptic augmentation in a cortical circuit model reproduces serial dependence in visual working memory. *Plos One* 12(12), p. e0188927.
- Bliss, D.P., Sun, J.J. and D’Esposito, M. 2017. Serial dependence is absent at the time of perception but increases in visual working memory. *Scientific Reports* 7(1), p. 14739.
- Bliss, T.V. and Collingridge, G.L. 1993. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* 361(6407), pp. 31–39.
- Brutkowski, S. 1965. Functions of prefrontal cortex in animals. *Physiological Reviews* 45(4), pp. 721–746.
- Buonomano, D.V. 1999. Distinct functional types of associative long-term potentiation in neocortical and hippocampal pyramidal neurons. *The Journal of Neuroscience* 19(16), pp. 6748–6754.
- Cano-Colino, M., Almeida, R., Gomez-Cabrero, D., Artigas, F. and Compte, A. 2014. Serotonin regulates performance nonmonotonically in a spatial working memory network. *Cerebral Cortex* 24(9), pp. 2449–2463.
- Cano-Colino, M. and Compte, A. 2012. A computational model for spatial working memory deficits in schizophrenia. *Pharmacopsychiatry* 45 Suppl 1, pp. S49-56.
- Carter, E. and Wang, X.-J. 2007. Cannabinoid-mediated disinhibition and working memory: dynamical interplay of multiple feedback mechanisms in a continuous attractor model of prefrontal cortex. *Cerebral Cortex* 17 Suppl 1, pp. i16-26.

- Castillo, P.E., Malenka, R.C. and Nicoll, R.A. 1997. Kainate receptors mediate a slow postsynaptic current in hippocampal CA3 neurons. *Nature* 388(6638), pp. 182–186.
- Castro-Alamancos, M.A. and Connors, B.W. 1996. Short-term synaptic enhancement and long-term potentiation in neocortex. *Proceedings of the National Academy of Sciences of the United States of America* 93(3), pp. 1335–1339.
- Castro-Alamancos, M.A., Donoghue, J.P. and Connors, B.W. 1995. Different forms of synaptic plasticity in somatosensory and motor areas of the neocortex. *The Journal of Neuroscience* 15(7 Pt 2), pp. 5324–5333.
- Catterall, W.A., Leal, K. and Nanou, E. 2013. Calcium channels and short-term synaptic plasticity. *The Journal of Biological Chemistry* 288(15), pp. 10742–10749.
- Catts, V.S., Lai, Y.L., Weickert, C.S., Weickert, T.W. and Catts, S.V. 2016. A quantitative review of the postmortem evidence for decreased cortical N-methyl-D-aspartate receptor expression levels in schizophrenia: How can we link molecular abnormalities to mismatch negativity deficits? *Biological Psychology* 116, pp. 57–67.
- Cicchini, G.M., Anobile, G. and Burr, D.C. 2014. Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform. *Proceedings of the National Academy of Sciences of the United States of America* 111(21), pp. 7867–7872.
- Cicchini, G.M., Mikellidou, K. and Burr, D.C. 2018. The functional role of serial dependence. *Proceedings. Biological Sciences / the Royal Society* 285(1890).
- Clopath, C. 2012. Synaptic consolidation: an approach to long-term learning. *Cognitive neurodynamics* 6(3), pp. 251–257.
- Clopath, C., Ziegler, L., Vasilaki, E., Büsing, L. and Gerstner, W. 2008. Tag-trigger-consolidation: a model of early and late long-term-potential and depression. *PLoS Computational Biology* 4(12), p. e1000248.
- Compte, A. 2006. Computational and in vitro studies of persistent activity: edging towards cellular and synaptic mechanisms of working memory. *Neuroscience* 139(1), pp. 135–151.
- Compte, A., Brunel, N., Goldman-Rakic, P.S. and Wang, X.J. 2000. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex* 10(9), pp. 910–923.
- Constantinidis, C., Franowicz, M.N. and Goldman-Rakic, P.S. 2001. Coding specificity in cortical microcircuits: a multiple-electrode analysis of primate prefrontal cortex. *The Journal of Neuroscience* 21(10), pp. 3646–3655.
- Conway, A.R.A., Kane, M.J. and Engle, R.W. 2003. Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences* 7(12), pp. 547–552.
- Corlew, R., Brasier, D.J., Feldman, D.E. and Philpot, B.D. 2008. Presynaptic NMDA receptors: newly appreciated roles in cortical synaptic function and plasticity. *The Neuroscientist* 14(6), pp. 609–625.
- Crair, M.C. and Malenka, R.C. 1995. A critical period for long-term potentiation at thalamocortical synapses. *Nature* 375(6529), pp. 325–328.
- Cull-Candy, S., Brickley, S. and Farrant, M. 2001. NMDA receptor subunits: diversity,

- development and disease. *Current Opinion in Neurobiology* 11(3), pp. 327–335.
- Dalmau, J., Gleichman, A.J., Hughes, E.G., et al. 2008. Anti-NMDA-receptor encephalitis: case series and analysis of the effects of antibodies. *Lancet Neurology* 7(12), pp. 1091–1098.
- Dalmau, J., Lancaster, E., Martinez-Hernandez, E., Rosenfeld, M.R. and Balice-Gordon, R. 2011. Clinical experience and laboratory investigations in patients with anti-NMDAR encephalitis. *Lancet Neurology* 10(1), pp. 63–74.
- Dalmau, J., Tüzün, E., Wu, H., et al. 2007. Paraneoplastic anti-N-methyl-D-aspartate receptor encephalitis associated with ovarian teratoma. *Annals of Neurology* 61(1), pp. 25–36.
- Dingledine, R., Borges, K., Bowie, D. and Traynelis, S.F. 1999. The glutamate receptor ion channels. *Pharmacological Reviews* 51(1), pp. 7–61.
- Driesen, N.R., Leung, H.-C., Calhoun, V.D., et al. 2008. Impairment of working memory maintenance and response in schizophrenia: functional magnetic resonance imaging evidence. *Biological Psychiatry* 64(12), pp. 1026–1034.
- Driesen, N.R., McCarthy, G., Bhagwagar, Z., et al. 2013. The impact of NMDA receptor blockade on human working memory-related prefrontal function and connectivity. *Neuropsychopharmacology* 38(13), pp. 2613–2622.
- Dudek, S.M. and Bear, M.F. 1993. Bidirectional long-term modification of synaptic effectiveness in the adult and immature hippocampus. *The Journal of Neuroscience* 13(7), pp. 2910–2918.
- Erickson, M.A., Maramba, L.A. and Lisman, J. 2010. A single brief burst induces GluR1-dependent associative short-term potentiation: a potential mechanism for short-term memory. *Journal of Cognitive Neuroscience* 22(11), pp. 2530–2540.
- Feldman, D.E. 2009. Synaptic mechanisms for plasticity in neocortex. *Annual Review of Neuroscience* 32, pp. 33–55.
- Fiebig, F. and Lansner, A. 2017. A Spiking Working Memory Model Based on Hebbian Short-Term Potentiation. *Journal of Neuroscience*, 37(1), pp. 83–96.
- Finke, C., Kopp, U.A., Prüss, H., Dalmau, J., Wandinger, K.-P. and Ploner, C.J. 2012. Cognitive deficits following anti-NMDA receptor encephalitis. *Journal of Neurology, Neurosurgery, and Psychiatry* 83(2), pp. 195–198.
- Fischer, J. and Whitney, D. 2014. Serial dependence in visual perception. *Nature Neuroscience* 17(5), pp. 738–743.
- Forbes, N.F., Carrick, L.A., McIntosh, A.M. and Lawrie, S.M. 2009. Working memory in schizophrenia: a meta-analysis. *Psychological Medicine* 39(6), pp. 889–905.
- Foster, J.J. and Awh, E. 2019. The role of alpha oscillations in spatial attention: limited evidence for a suppression account. *Current opinion in psychology* 29, pp. 34–40.
- Foster, J.J., Sutterer, D.W., Serences, J.T., Vogel, E.K. and Awh, E. 2016. The topography of alpha-band activity tracks the content of spatial working memory. *Journal of Neurophysiology* 115(1), pp. 168–177.

- Fritsche, M., Mostert, P. and de Lange, F.P. 2017. Opposite effects of recent history on perception and decision. *Current Biology* 27(4), pp. 590–595.
- Fujisawa, S., Amarasingham, A., Harrison, M.T. and Buzsáki, G. 2008. Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nature Neuroscience* 11(7), pp. 823–833.
- Funahashi, S., Bruce, C.J. and Goldman-Rakic, P.S. 1989. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology* 61(2), pp. 331–349.
- Fuster, J.M. 1973. Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. *Journal of Neurophysiology* 36(1), pp. 61–78.
- Fuster, J.M. and Alexander, G.E. 1973. Firing changes in cells of the nucleus medialis dorsalis associated with delayed response behavior. *Brain Research* 61, pp. 79–91.
- Fuster, J.M. and Alexander, G.E. 1971. Neuron activity related to short-term memory. *Science* 173(3997), pp. 652–654.
- Gao, Z., Davis, C., Thomas, A.M., et al. 2018. A cortico-cerebellar loop for motor planning. *Nature* 563(7729), pp. 113–116.
- Glantz, L.A. and Lewis, D.A. 2000. Decreased dendritic spine density on prefrontal cortical pyramidal neurons in schizophrenia. *Archives of General Psychiatry* 57(1), pp. 65–73.
- Goldman-Rakic, P.S. 1995. Cellular basis of working memory. *Neuron* 14(3), pp. 477–485.
- Goldman-Rakic, P.S. 1987. Development of cortical circuitry and cognitive function. *Child Development* 58(3), pp. 601–622.
- Gold, J.M., Bansal, S., Anticevic, A., et al. 2020. Refining the empirical constraints on computational models of spatial working memory in schizophrenia. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 5(9), pp. 913–922.
- Gold, J.M., Hahn, B., Zhang, W.W., et al. 2010. Reduced capacity but spared precision and maintenance of working memory representations in schizophrenia. *Archives of General Psychiatry* 67(6), pp. 570–577.
- González-Burgos, G., Barrionuevo, G. and Lewis, D.A. 2000. Horizontal synaptic connections in monkey prefrontal cortex: an in vitro electrophysiological study. *Cerebral Cortex* 10(1), pp. 82–92.
- Gorgoraptis, N., Catalao, R.F.G., Bays, P.M. and Husain, M. 2011. Dynamic updating of working memory resources for visual objects. *The Journal of Neuroscience* 31(23), pp. 8502–8511.
- Gould, I.C., Rushworth, M.F. and Nobre, A.C. 2011. Indexing the graded allocation of visuospatial attention using anticipatory alpha oscillations. *Journal of Neurophysiology* 105(3), pp. 1318–1326.
- Grech, R., Cassar, T., Muscat, J., et al. 2008. Review on solving the inverse problem in EEG source analysis. *Journal of Neuroengineering and Rehabilitation* 5, p. 25.
- Hahn, B., Robinson, B.M., Leonard, C.J., Luck, S.J. and Gold, J.M. 2018. Posterior parietal cortex dysfunction is central to working memory storage and broad cognitive deficits in

- schizophrenia. *The Journal of Neuroscience* 38(39), pp. 8378–8387.
- Haj-Dahmane, S. and Andrade, R. 1998. Ionic mechanism of the slow afterdepolarization induced by muscarinic receptor activation in rat prefrontal cortex. *Journal of Neurophysiology* 80(3), pp. 1197–1210.
- Hansel, D. and Mato, G. 2013. Short-term plasticity explains irregular persistent activity in working memory tasks. *The Journal of Neuroscience* 33(1), pp. 133–149.
- Harrison, P.J. and Weinberger, D.R. 2005. Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. *Molecular Psychiatry* 10(1), pp. 40–68; image 5.
- Hebb, D.O. 1949. *The organization of behavior*. New York: Wiley.
- Heeger, D.J. and Ress, D. 2002. What does fMRI tell us about neuronal activity? *Nature Reviews. Neuroscience* 3(2), pp. 142–151.
- Hempel, C.M., Hartman, K.H., Wang, X.J., Turrigiano, G.G. and Nelson, S.B. 2000. Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex. *Journal of Neurophysiology* 83(5), pp. 3031–3041.
- Hemsley, D.R. 2005. The development of a cognitive model of schizophrenia: placing it in context. *Neuroscience and Biobehavioral Reviews* 29(6), pp. 977–988.
- Hermoso-Mendizabal, A., Hyafil, A., Rueda-Orozco, P.E., Jaramillo, S., Robbe, D. and de la Rocha, J. 2020. Response outcomes gate the impact of expectations on perceptual decisions. *Nature Communications* 11(1), p. 1057.
- Herwig, A., Beisert, M. and Schneider, W.X. 2010. On the spatial interaction of visual working memory and attention: evidence for a global effect from memory-guided saccades. *Journal of Vision* 10(5), p. 8.
- Hirsch, J.C. and Crepel, F. 1990. Use-dependent changes in synaptic efficacy in rat prefrontal neurons in vitro. *The Journal of Physiology* 427, pp. 31–49.
- Honey, R.A.E., Honey, G.D., O’Loughlin, C., et al. 2004. Acute ketamine administration alters the brain responses to executive demands in a verbal working memory task: an FMRI study. *Neuropsychopharmacology* 29(6), pp. 1203–1214.
- Hubel, D.H. and Wiesel, T.N. 1959. Receptive fields of single neurones in the cat’s striate cortex. *The Journal of Physiology* 148, pp. 574–591.
- Hughes, E.G., Peng, X., Gleichman, A.J., et al. 2010. Cellular and synaptic mechanisms of anti-NMDA receptor encephalitis. *The Journal of Neuroscience* 30(17), pp. 5866–5875.
- Hunter, W.S. 1913. The Delayed Reaction in Animals and Children. *Animal Behavior Monographs* 2(1), p. 86.
- Inagaki, H.K., Fontolan, L., Romani, S. and Svoboda, K. 2019. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature* 566(7743), pp. 212–217.
- Itskov, V., Hansel, D. and Tsodyks, M. 2011. Short-Term Facilitation may Stabilize Parametric Working Memory Trace. *Frontiers in Computational Neuroscience* 5, p. 40.
- Kayser, M.S. and Dalmau, J. 2014. Anti-NMDA receptor encephalitis, autoimmunity, and

psychosis. *Schizophrenia Research* 176(1), pp. 36–40.

Kelly, S.P., Lalor, E.C., Reilly, R.B. and Foxe, J.J. 2006. Increases in alpha oscillatory power reflect an active retinotopic mechanism for distracter suppression during sustained visuospatial attention. *Journal of Neurophysiology* 95(6), pp. 3844–3851.

Kilpatrick, Z.P. 2018. Synaptic mechanisms of interference in working memory. *Scientific Reports* 8(1), p. 7879.

Kim, S.S., Rouault, H., Druckmann, S. and Jayaraman, V. 2017. Ring attractor dynamics in the *Drosophila* central brain. *Science* 356(6340), pp. 849–853.

King, J.R. and Dehaene, S. 2014. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences* 18(4), pp. 203–210.

Kirkwood, A. and Bear, M.F. 1994. Hebbian synapses in visual cortex. *The Journal of Neuroscience* 14(3 Pt 2), pp. 1634–1645.

Kiyonaga, A., Scimeca, J.M., Bliss, D.P. and Whitney, D. 2017. Serial Dependence across Perception, Attention, and Memory. *Trends in Cognitive Sciences* 21(7), pp. 493–497.

Kojima, S. and Goldman-Rakic, P.S. 1982. Delay-related activity of prefrontal neurons in rhesus monkeys performing delayed response. *Brain Research* 248(1), pp. 43–49.

Kristiansen, L.V., Beneyto, M., Haroutunian, V. and Meador-Woodruff, J.H. 2006. Changes in NMDA receptor subunits and interacting PSD proteins in dorsolateral prefrontal and anterior cingulate cortex indicate abnormal regional expression in schizophrenia. *Molecular Psychiatry* 11(8), pp. 737–47, 705.

Kritzer, M.F. and Goldman-Rakic, P.S. 1995. Intrinsic circuit organization of the major layers and sublayers of the dorsolateral prefrontal cortex in the rhesus monkey. *The Journal of Comparative Neurology* 359(1), pp. 131–143.

Krystal, J.H., Anticevic, A., Yang, G.J., et al. 2017. Impaired tuning of neural ensembles and the pathophysiology of schizophrenia: A translational and computational neuroscience perspective. *Biological Psychiatry* 81(10), pp. 874–885.

Kubota, K. and Niki, H. 1971. Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology* 34(3), pp. 337–347.

Leavitt, M.L., Mendoza-Halliday, D. and Martinez-Trujillo, J.C. 2017. Sustained activity encoding working memories: not fully distributed. *Trends in Neurosciences* 40(6), pp. 328–346.

Lebovich, L., Darshan, R., Lavi, Y., Hansel, D. and Loewenstein, Y. 2019. Idiosyncratic choice bias naturally emerges from intrinsic stochasticity in neuronal dynamics. *Nature human behaviour* 3(11), pp. 1190–1202.

Lee, J. and Park, S. 2005. Working memory impairments in schizophrenia: a meta-analysis. *Journal of Abnormal Psychology* 114(4), pp. 599–611.

Lennox, B.R., Coles, A.J. and Vincent, A. 2012. Antibody-mediated encephalitis: a treatable cause of schizophrenia. *The British Journal of Psychiatry* 200(2), pp. 92–94.

Leonard, C.J., Robinson, B.M., Hahn, B., Luck, S.J. and Gold, J.M. 2017. Altered spatial profile of distraction in people with schizophrenia. *Journal of Abnormal Psychology* 126(8),

pp. 1077–1086.

Lester, R.A., Clements, J.D., Westbrook, G.L. and Jahr, C.E. 1990. Channel kinetics determine the time course of NMDA receptor-mediated synaptic currents. *Nature* 346(6284), pp. 565–567.

Lewis, D.A., Hashimoto, T. and Volk, D.W. 2005. Cortical inhibitory neurons and schizophrenia. *Nature Reviews. Neuroscience* 6(4), pp. 312–324.

Liang, S.G. and Greenwood, T.A. 2015. The impact of clinical heterogeneity in schizophrenia on genomic analyses. *Schizophrenia Research* 161(2–3), pp. 490–495.

Lieberman, A., Fischer, J. and Whitney, D. 2014. Serial dependence in the perception of faces. *Current Biology* 24(21), pp. 2569–2574.

Lieder, I., Adam, V., Frenkel, O., Jaffe-Dax, S., Sahani, M. and Ahissar, M. 2019. Perceptual bias reveals slow-updating in autism and fast-forgetting in dyslexia. *Nature Neuroscience* 22(2), pp. 256–264.

Lim, S. and Goldman, M.S. 2013. Balanced cortical microcircuitry for maintaining information in working memory. *Nature Neuroscience* 16(9), pp. 1306–1314.

Lipinski, J., Simmering, V.R., Johnson, J.S. and Spencer, J.P. 2010. The role of experience in location estimation: Target distributions shift location memory biases. *Cognition* 115(1), pp. 147–153.

Lisman, J.E., Fellous, J.M. and Wang, X.J. 1998. A role for NMDA-receptor channels in working memory. *Nature Neuroscience* 1(4), pp. 273–275.

Liu, D., Gu, X., Zhu, J., et al. 2014. Medial prefrontal activity during delay period contributes to learning of a working memory task. *Science* 346(6208), pp. 458–463.

Lorente De Nó, R. 1938. Analysis of the activity of the chains of internuncial neurons. *Journal of neurophysiology* 1(3), pp. 207–244.

Lundqvist, M., Herman, P. and Miller, E.K. 2018. Working memory: delay activity, yes! persistent activity? maybe not. *The Journal of Neuroscience* 38(32), pp. 7013–7019.

Mäki-Marttunen, T., Iannella, N., Edwards, A.G., Einevoll, G.T. and Blackwell, K.T. 2020. A unified computational model for cortical post-synaptic plasticity. *eLife* 9.

Malenka, R.C. 2002. Synaptic Plasticity. In: Davis, K. L., Charney, D., Coyle, J. T., and Nemeroff, C. eds. *Neuropsychopharmacology – 5th Generation of Progress*.

Malenka, R.C. and Bear, M.F. 2004. LTP and LTD: an embarrassment of riches. *Neuron* 44(1), pp. 5–21.

Malenka, R.C. and Nicoll, R.A. 1999. Long-term potentiation - a decade of progress? *Science* 285(5435), pp. 1870–1874.

Maneta, E. and Garcia, G. 2014. Psychiatric manifestations of anti-NMDA receptor encephalitis: neurobiological underpinnings and differential diagnostic implications. *Psychosomatics* 55(1), pp. 37–44.

Marder, E. 2012. Neuromodulation of neuronal circuits: back to the future. *Neuron* 76(1), pp. 1–11.

- Marder, E., Abbott, L.F., Turrigiano, G.G., Liu, Z. and Golowasch, J. 1996. Memory from the dynamics of intrinsic membrane currents. *Proceedings of the National Academy of Sciences of the United States of America* 93(24), pp. 13481–13486.
- Marder, E. and Goaillard, J.-M. 2006. Variability, compensation and homeostasis in neuron and network function. *Nature Reviews. Neuroscience* 7(7), pp. 563–574.
- Marder, E. and Taylor, A.L. 2011. Multiple models to capture the variability in biological neurons and networks. *Nature Neuroscience* 14(2), pp. 133–138.
- Markram, H., Wang, Y., and Tsodyks, M. 1998. Differential signaling via the same axon of neocortical pyramidal neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 95(9), pp. 5323-5328.
- Masdeu, J.C., Dalmau, J. and Berman, K.F. 2016. NMDA receptor internalization by autoantibodies: A reversible mechanism underlying psychosis? *Trends in Neurosciences* 39(5), pp. 300–310.
- McKeon, G.L., Robinson, G.A., Ryan, A.E., et al. 2018. Cognitive outcomes following anti-N-methyl-D-aspartate receptor encephalitis: A systematic review. *Journal of Clinical and Experimental Neuropsychology* 40(3), pp. 234–252.
- Mejias, J.F. and Wang, X.-J. 2019. Mechanisms of distributed working memory in a large-scale model of the macaque neocortex. *BioRxiv*.
- Mendoza-Halliday, D. and Martinez-Trujillo, J.C. 2017. Neuronal population coding of perceived and memorized visual features in the lateral prefrontal cortex. *Nature Communications* 8, p. 15471.
- Meunier, C.N.J., Chameau, P. and Fossier, P.M. 2017. Modulation of synaptic plasticity in the cortex needs to understand all the players. *Frontiers in synaptic neuroscience* 9, p. 2.
- Meyers, E.M., Freedman, D.J., Kreiman, G., Miller, E.K. and Poggio, T. 2008. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of Neurophysiology* 100(3), pp. 1407–1419.
- Michel, C.M. and Brunet, D. 2019. EEG source imaging: A practical review of the analysis steps. *Frontiers in neurology* 10, p. 325.
- Miller, E.K., Erickson, C.A. and Desimone, R. 1996. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *The Journal of Neuroscience* 16(16), pp. 5154–5167.
- Miller, E.K., Lundqvist, M. and Bastos, A.M. 2018. Working Memory 2.0. *Neuron* 100(2), pp. 463–475.
- Miller, G.A. 1956. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63(2), pp. 81–97.
- Mochol, G., Kiani, R. and Moreno-Bote, R. 2020. Prefrontal cortex represents heuristics that shape choice bias and its integration into future behavior. *BioRxiv*.
- Mongillo, G., Barak, O. and Tsodyks, M. 2008. Synaptic theory of working memory. *Science* 319(5869), pp. 1543–1546.
- Mongillo, G., Hansel, D. and van Vreeswijk, C. 2012. Bistability and spatiotemporal

irregularity in neuronal networks with nonlinear synaptic transmission. *Physical Review Letters* 108(15), p. 158101.

Morgan, C.J.A. and Curran, H.V. 2006. Acute and chronic effects of ketamine upon human memory: a review. *Psychopharmacology* 188(4), pp. 408–424.

Morris, R.G. 1989. Synaptic plasticity and learning: selective impairment of learning rats and blockade of long-term potentiation in vivo by the N-methyl-D-aspartate receptor antagonist AP5. *The Journal of Neuroscience* 9(9), pp. 3040–3057.

Murray, J.D., Anticevic, A., Gancsos, M., et al. 2014. Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. *Cerebral Cortex* 24(4), pp. 859–872.

Murray, J.D., Bernacchia, A., Freedman, D.J., et al. 2014. A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience* 17(12), pp. 1661–1663.

Murray, J.D., Bernacchia, A., Roy, N.A., Constantinidis, C., Romo, R. and Wang, X.-J. 2017. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America* 114(2), pp. 394–399.

Nanou, E. and Catterall, W.A. 2018. Calcium channels, synaptic plasticity, and neuropsychiatric disease. *Neuron* 98(3), pp. 466–481.

Nassar, M.R., Helmers, J.C. and Frank, M.J. 2018. Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological Review* 125(4), pp. 486–511.

Niki, H. 1974. Differential activity of prefrontal units during right and left delayed response trials. *Brain Research* 70(2), pp. 346–349.

Niki, H. and Watanabe, M. 1976. Prefrontal unit activity and delayed response: relation to cue location versus direction of response. *Brain Research* 105(1), pp. 79–88.

Nowak, L., Bregestovski, P., Ascher, P., Herbet, A. and Prochiantz, A. 1984. Magnesium gates glutamate-activated channels in mouse central neurones. *Nature* 307(5950), pp. 462–465.

Oh, B.-I., Kim, Y.-J. and Kang, M.-S. 2019. Ensemble representations reveal distinct neural coding of visual working memory. *Nature Communications* 10(1), p. 5665.

Olney, J.W., Newcomer, J.W. and Farber, N.B. 1999. NMDA receptor hypofunction model of schizophrenia. *Journal of Psychiatric Research* 33(6), pp. 523–533.

Orhan, A.E. and Ma, W.J. 2019. A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nature Neuroscience* 22(2), pp. 275–283.

Oviedo-Salcedo, T., de Witte, L., Kümpfel, T., et al. 2018. Absence of cerebrospinal fluid antineuronal antibodies in schizophrenia spectrum disorders. *The British Journal of Psychiatry* 212(5), pp. 318–320.

Panichello, M.F. and Buschman, T.J. 2020. Selective control of working memory in prefrontal, parietal, and visual cortex. *BioRxiv*.

Panichello, M.F., DePasquale, B., Pillow, J.W. and Buschman, T.J. 2019. Error-correcting

- dynamics in visual working memory. *Nature Communications* 10(1), p. 3366.
- Papadimitriou, C., Ferdoash, A. and Snyder, L.H. 2015. Ghosts in the machine: memory interference from the previous trial. *Journal of Neurophysiology* 113(2), pp. 567–577.
- Papadimitriou, C., White, R.L. and Snyder, L.H. 2017. Ghosts in the Machine II: Neural Correlates of Memory Interference from the Previous Trial. *Cerebral Cortex* 27(4), pp. 2513–2527.
- Park, S. and Holzman, P.S. 1992. Schizophrenics show spatial working memory deficits. *Archives of General Psychiatry* 49(12), pp. 975–982.
- Parthasarathy, A., Tang, C., Herikstad, R., Cheong, L.F., Yen, S.-C. and Libedinsky, C. 2019. Time-invariant working memory representations in the presence of code-morphing in the lateral prefrontal cortex. *Nature Communications* 10(1), p. 4995.
- Pasternak, T. and Greenlee, M.W. 2005. Working memory in primate sensory systems. *Nature Reviews. Neuroscience* 6(2), pp. 97–107.
- Pelletier, M., Achim, A.M., Montoya, A., Lal, S. and Lepage, M. 2005. Cognitive and clinical moderators of recognition memory in schizophrenia: a meta-analysis. *Schizophrenia Research* 74(2–3), pp. 233–252.
- Pereira, J. and Wang, X.-J. 2015. A Tradeoff Between Accuracy and Flexibility in a Working Memory Circuit Endowed with Slow Feedback Mechanisms. *Cerebral Cortex* 25(10), pp. 3586–3601.
- Planagumà, J., Haselmann, H., Mannara, F., et al. 2016. Ephrin-B2 prevents N-methyl-D-aspartate receptor antibody effects on memory and neuroplasticity. *Annals of Neurology* 80(3), pp. 388–400.
- Planagumà, J., Leyboldt, F., Mannara, F., et al. 2015. Human N-methyl D-aspartate receptor antibodies alter memory and behaviour in mice. *Brain: A Journal of Neurology* 138(Pt 1), pp. 94–109.
- Rademaker, R.L., Bloem, I.M., De Weerd, P. and Sack, A.T. 2015. The impact of interference on short-term memory for visual orientation. *Journal of Experimental Psychology. Human Perception and Performance* 41(6), pp. 1650–1665.
- Rademaker, R.L., Park, Y.E., Sack, A.T. and Tong, F. 2018. Evidence of gradual loss of precision for simple features and complex objects in visual working memory. *Journal of Experimental Psychology. Human Perception and Performance* 44(6), pp. 925–940.
- Rao, S.G., Williams, G.V. and Goldman-Rakic, P.S. 1999. Isodirectional tuning of adjacent interneurons and pyramidal cells during working memory: evidence for microcolumnar organization in PFC. *Journal of Neurophysiology* 81(4), pp. 1903–1916.
- Reinhart, R.M.G., Heitz, R.P., Purcell, B.A., Weigand, P.K., Schall, J.D. and Woodman, G.F. 2012. Homologous mechanisms of visuospatial working memory maintenance in macaque and human: properties and sources. *The Journal of Neuroscience* 32(22), pp. 7711–7722.
- Rihs, T.A., Michel, C.M. and Thut, G. 2007. Mechanisms of selective inhibition in visual spatial attention are indexed by alpha-band EEG synchronization. *The European Journal of Neuroscience* 25(2), pp. 603–610.

- Riley, M.R. and Constantinidis, C. 2015. Role of Prefrontal Persistent Activity in Working Memory. *Frontiers in Systems Neuroscience* 9, p. 181.
- Rolls, E.T., Loh, M., Deco, G. and Winterer, G. 2008. Computational models of schizophrenia and dopamine modulation in the prefrontal cortex. *Nature Reviews. Neuroscience* 9(9), pp. 696–709.
- Rose, N.S., LaRocque, J.J., Riggall, A.C., et al. 2016. Reactivation of latent working memories with transcranial magnetic stimulation. *Science* 354(6316), pp. 1136–1139.
- Sakai, K. and Passingham, R.E. 2003. Prefrontal interactions reflect future task operations. *Nature Neuroscience* 6(1), pp. 75–81.
- Scoville, W.B. and Milner, B. 1957. Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry* 20(1), pp. 11–21.
- Seeholzer, A., Deger, M. and Gerstner, W. 2019. Stability of working memory in continuous attractor networks under the control of short-term plasticity. *PLoS Computational Biology* 15(4), p. e1006928.
- Shafi, M., Zhou, Y., Quintana, J., Chow, C., Fuster, J. and Bodner, M. 2007. Variability in neuronal activity in primate cortex during working memory tasks. *Neuroscience* 146(3), pp. 1082–1108.
- Shallice, T. and Warrington, E.K. 1970. Independent functioning of verbal memory stores: a neuropsychological study. *The Quarterly journal of experimental psychology* 22(2), pp. 261–273.
- Shih, P.-Y., Savtchenko, L.P., Kamasawa, N., et al. 2013. Retrograde synaptic signaling mediated by K⁺ efflux through postsynaptic NMDA receptors. *Cell reports* 5(4), pp. 941–951.
- Shin, H., Zou, Q. and Ma, W.J. 2017. The effects of delay duration on visual working memory for orientation. *Journal of Vision* 17(14), p. 10.
- Sitskoorn, M.M., Aleman, A., Ebisch, S.J.H., Appels, M.C.M. and Kahn, R.S. 2004. Cognitive deficits in relatives of patients with schizophrenia: a meta-analysis. *Schizophrenia Research* 71(2–3), pp. 285–295.
- Snyder, A.C. and Smith, M.A. 2015. Stimulus-dependent spiking relationships with the EEG. *Journal of Neurophysiology* 114(3), pp. 1468–1482.
- Sohal, V.S. and Rubenstein, J.L.R. 2019. Excitation-inhibition balance as a framework for investigating mechanisms in neuropsychiatric disorders. *Molecular Psychiatry* 24(9), pp. 1248–1257.
- Spaak, E., Watanabe, K., Funahashi, S. and Stokes, M.G. 2017. Stable and dynamic coding for working memory in primate prefrontal cortex. *The Journal of Neuroscience* 37(27), pp. 6503–6516.
- Starc, M., Murray, J.D., Santamauro, N., et al. 2017. Schizophrenia is associated with a pattern of spatial working memory deficits consistent with cortical disinhibition. *Schizophrenia Research* 181, pp. 107–116.
- Steiner, J., Walter, M., Glanz, W., et al. 2013. Increased prevalence of diverse

N-methyl-D-aspartate glutamate receptor antibodies in patients with an initial diagnosis of schizophrenia: specific relevance of IgG NR1a antibodies for distinction from N-methyl-D-aspartate glutamate receptor encephalitis. *JAMA psychiatry* 70(3), pp. 271–278.

Stephan, K.E., Baldeweg, T. and Friston, K.J. 2006. Synaptic plasticity and dysconnection in schizophrenia. *Biological Psychiatry* 59(10), pp. 929–939.

Stokes, M.G. 2015. “Activity-silent” working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences* 19(7), pp. 394–405.

Stokes, M.G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D. and Duncan, J. 2013. Dynamic coding for cognitive control in prefrontal cortex. *Neuron* 78(2), pp. 364–375.

Suzuki, M. and Gottlieb, J. 2013. Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nature Neuroscience* 16(1), pp. 98–104.

Tsodyks, M. and Markram, H. 1997. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proceedings of the National Academy of Sciences of the United States of America* 94(2), pp. 719–723.

Van der Stigchel, S., Merten, H., Meeter, M. and Theeuwes, J. 2007. The effects of a task-irrelevant visual event on spatial working memory. *Psychonomic Bulletin & Review* 14(6), pp. 1066–1071.

Volianskis, A., Bannister, N., Collett, V.J., et al. 2013. Different NMDA receptor subtypes mediate induction of long-term potentiation and two forms of short-term potentiation at CA1 synapses in rat hippocampus in vitro. *The Journal of Physiology* 591(4), pp. 955–972.

Volianskis, A., France, G., Jensen, M.S., Bortolotto, Z.A., Jane, D.E. and Collingridge, G.L. 2015. Long-term potentiation and the role of N-methyl-D-aspartate receptors. *Brain Research* 1621, pp. 5–16.

Wang, H., Stradtman, G.G., Wang, X.-J. and Gao, W.-J. 2008. A specialized NMDA receptor function in layer 5 recurrent microcircuitry of the adult rat prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America* 105(43), pp. 16791–16796.

Wang, M., Yang, Y., Wang, C.-J., et al. 2013. NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. *Neuron* 77(4), pp. 736–749.

Wang, X.J. 1999. Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *The Journal of Neuroscience* 19(21), pp. 9587–9603.

Wang, X.J. 2001. Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neurosciences* 24(8), pp. 455–463.

Wang, Y., Markram, H., Goodman, P.H., Berger, T.K., Ma, J. and Goldman-Rakic, P.S. 2006. Heterogeneity in the pyramidal network of the medial prefrontal cortex. *Nature Neuroscience* 9(4), pp. 534–542.

Weickert, C.S. and Weickert, T.W. 2016. What’s hot in schizophrenia research? *The Psychiatric Clinics of North America* 39(2), pp. 343–351.

Wilken, P. and Ma, W.J. 2004. A detection theory account of change detection. *Journal of*

Vision 4(12), pp. 1120–1135.

Wimmer, K., Nykamp, D.Q., Constantinidis, C. and Compte, A. 2014. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience* 17(3), pp. 431–439.

Wolff, M.J., Jochim, J., Akyürek, E.G., Buschman, T.J. and Stokes, M.G. 2020. Drifting codes within a stable coding scheme for working memory. *PLoS Biology* 18(3), p. e3000625.

Wolff, M.J., Jochim, J., Akyürek, E.G. and Stokes, M.G. 2017. Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience* 20(6), pp. 864–871.

Xu, Y. 2017. Reevaluating the sensory account of visual working memory storage. *Trends in Cognitive Sciences* 21(10), pp. 794–815.

Yang, G.J., Murray, J.D., Wang, X.-J., et al. 2016. Functional hierarchy underlies preferential connectivity disturbances in schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America* 113(2), pp. E219-28.

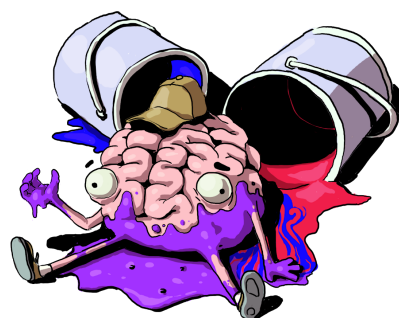
Yoon, J.Y., Lee, H.R., Ho, W.-K. and Lee, S.-H. 2020. Disparities in Short-Term Depression Among Prefrontal Cortex Synapses Sustain Persistent Activity in a Balanced Network. *Cerebral Cortex* 30(1), pp. 113–134.

York, L.C. and van Rossum, M.C.W. 2009. Recurrent networks with short term synaptic depression. *Journal of Computational Neuroscience* 27(3), pp. 607–620.

Zhang, W. and Luck, S.J. 2008. Discrete fixed-resolution representations in visual working memory. *Nature* 453(7192), pp. 233–235.

Zucker, R.S. and Regehr, W.G. 2002. Short-term synaptic plasticity. *Annual Review of Physiology* 64, pp. 355–405.

Zylberberg, J. and Strowbridge, B.W. 2017. Mechanisms of persistent activity in cortical circuits: possible neural substrates for working memory. *Annual Review of Neuroscience* 40, pp. 603–627.



© Pedro Podre