UNIVERSITAT DE BARCELONA

Circuit mechanisms of working memory maintenance and distractor interference

David Sánchez Bestué

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (**www.tdx.cat**) i a través del Dipòsit Digital de la UB (**diposit.ub.edu**) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (**www.tdx.cat**) y a través del Repositorio Digital de la UB (**diposit.ub.edu**) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (**www.tdx.cat**) service and by the UB Digital Repository (**diposit.ub.edu**) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

UNIVERSITAT DE BARCELONA

Programa de Doctorat en Biomedicina

Circuit mechanisms of working memory maintenance and distractor interference

David Sánchez Bestué

Supervised by: Albert Compte and Rita Almeida

Barcelona, 2021

UNIVERSITAT DE BARCELONA

Programa de Doctorat en Biomedicina

Area: Neuroscience

Line: Neurophysiology and computation in cortical systems

Circuit mechanisms of working memory maintenance and distractor interference

David Sánchez Bestué

Supervisor IDIBAPS:
Albert Compte Braquets

Supervisor Karolinska Institutet and Stockholm University:
Rita Almeida

Tutor Universitat de Barcelona:
Josep Dalmau Obrador

Barcelona, 2021

“Don’t move until you see it”

- Searching for Bobby Fischer-

Abstract

The main goal of this thesis is to study the circuit for working memory maintenance from a mechanistic perspective. To do so, I combine behavioral experiments with neuroimaging techniques and neuronal recordings under the framework of the bump attractor model for visuospatial working memory. The work performed during this thesis is encapsulated in two main chapters, one focusing on describing the topography of the working memory circuit, and the other focusing on how distractors interfere with the working memory content.

In the Chapter *Topography of the working memory circuit*, I test the Sensory Recruitment Theory by evaluating whether encoding and maintenance have the same topographical signatures, as expected if they share the same neural circuit. I provide behavioral, modeling, and electrophysiological data supporting the idea that prefrontal working memory maintenance is separated from encoding processes and mediated by attractor dynamics. Furthermore, I will extend the bump attractor model to cover the radial dimension and provide a mechanistic explanation for the compression of the visual field (foveal bias) with delay.

In the Chapter *Distractor filtering in the working memory circuit*, I evaluate how distractors interfere spatially and temporally with working memory maintenance at the behavioral and fMRI levels. I evaluate the results in the framework of the bump attractor model, and I explore different control strategies to deal with distracting information. Furthermore, in this chapter I re-analyze two electrophysiological datasets (Suzuki & Gottlieb, Nat. Neurosc., 2013 and Qi et al., Cell Reports, 2021) to test some predictions of the model and to mechanistically explain cholinergic improvement of working memory in a distractor-filtering task when stimulating the Nucleus Basalis of Meynert.

Table of contents

Acknowledgements

Si tuviese que escoger las dos frases que más han influido en mi vida, posiblemente serían estas: *"mayorcito para irse de fiesta, mayorcito para levantarse"* y *"Cristina, la vida és un teatre"*. La primera es una frase recurrente de mi padre, mientras que la segunda se la dijo mi profesora de catalán a una buena amiga para consolarla. Más que para hacer una disertación filosófica sobre el significado de ambas, las voy a utilizar para mostrar mi agradecimiento a dos pilares sin los que este trabajo -ni nada- habría sido posible: mi familia y mis mentores.

Siempre me he considerado un privilegiado por tener una familia unida. Poder confiar siempre en ellos, contar con su apoyo, su crítica y saber que, pase lo que pase, siempre me ayudarán, es un tesoro incalculable. Gracias a mis padres y a mi hermana por el ejemplo que me dais día a día, por dejarme crecer y por el amor que me hacéis sentir. Literalmente, la vida no tendría sentido sin vosotros. A su lado están mis abuelos, los que están y el que no, por las enseñanzas de vida que año a año entiendo más. La admiración que tengo por vosotros es infinita. Muchas gracias a mis primos y mis tíos, cercanos y lejanos, por demostrar que el cariño no se mide en las veces que nos vemos; y también a los nuevos miembros, como mi cuñado Pablo, que hacen que la familia siga creciendo.

Siempre he pensado que una tesis, aunque aborde un campo muy específico del conocimiento, se nutre de todos aquellos mentores que nos han influenciado a lo largo de la vida. Es inevitable volver a hablar de mis padres y de la educación que me han dado, así como de aquellos que plantaron la semilla de la curiosidad científica como mis queridos Toni Valls o Marta Segura, entre otros. Sin ellos, y sin todos aquellos maestros que me enseñaron tanto conocimiento como valores, esta tesis sería imposible.

Antes de hablar de mis supervisores directos, he de agradecer a aquellos que me introdujeron en el mundo de la neurociencia. Mi vida sería completamente distinta si, de camino a una cena de magos, hubiese ido en el coche principal en vez de bajar andando junto al gran mago y amigo Eduard Juanola. Comentándole que me gustaría estudiar el cerebro usando la magia como vehículo respondió: *"Que no coneixes a en Jordi*

Camí?". Me da vértigo pensar que sin ese paseo nada de esto hubiese pasado, porque de ahí conocí a Jordi, una de las personas más generosas que jamás conoceré y que desde el primer día me "apadrinó" y hasta el día de hoy me sigue ayudando en todo lo que puede. Gracias eternas, Jordi. Junto a esa figura, se encuentra otra de la misma envergadura: Luis Martínez. La experiencia que tan generosamente me dio Luis de ir a su laboratorio en el Instituto de Neurociencias de Alicante para hacer el TFG es una de las vivencias más transformadoras de mi vida y de la cual estoy increíblemente agradecido. Mil gracias, Luis, Sandra y Mari Carmen por todo lo que aprendí y crecí con vosotros. Siempre digo que en una semana allí aprendí más que en 4 años de universidad y el "primero el uno y luego el dos" o las correcciones que le hicieron al mismísimo Torsten Wiesel, sigue presente hasta el día de hoy.

Aunque de todas las cosas que le he de agradecer a Luis, sin duda la mayor de ellas fue la recomendación de que intentase hacer mi tesis del máster con un tal Albert Compte. Mil gracias, Albert por darme estas oportunidades, por apoyarme durante estos años y enseñarme tanto. Se dice que los mejores mentores son aquellos que, además de darte conocimiento, te motivan a encontrarlo. Por eso gracias de nuevo, por dejarme explorar con libertad y por ayudarme siempre que lo he necesitado.

De Barcelona salté a Estocolmo, y allí encontré a quien sería mi ángel de la guardia todo este tiempo. No tengo palabras suficientes para describir el agradecimiento que siento por ti, Rita, por haber estado, día a día, semana a semana, mes a mes y año a año, desde cerca y desde lejos cuidándome para que esta tesis vea la luz. Si algún día puedo devolverte una centésima parte de todo lo que me has dado con una centésima parte de la humildad, la cercanía y el cariño con lo que lo has hecho tú, me sentiría más que orgulloso. Aunque no puedo nombraros a todos, thanks Torkel for your hospitality and scientific determination, thanks Nick for being a remote friend, always taking care of keeping in touch. Besides being an outstanding and humble scientist, I have always found fascinating your ability to be loved by everyone. Thanks to all the lab members I have coincided with there, my "old best friend" and my big tennis rival Douglas, the fascinating Bruno, Jet, Sophie, Karen, Fahimeh, Federico, Linnea, Teng, George, Amy… and of course to my flatmates: Kjelle and Angus (and Linnea!), my handball guys: Eric, Christos, Niklas, Victor, Pontus, Macki…

and other good friends (Anna, Laura, Fernando, Pere, Elena, Ana, Belén, Stefan, Robert, Erik…) that made those two years so special.

También quiero agradecer a todos los miembros del laboratorio, desde la "vieja guardia" a las "nuevas generaciones". Gracias Joao por tener el honor de ser el primero de tu ejército de "master students", también a Genís, que no se va, pero al que tampoco dejaremos ir… a Ainoha, por ser la única cuerda entre tanto loco, a la "leyenda", a Max, a Marc, a Pablo… y también por supuesto a los "nuevos": a Dani Duque, mi socio, que me demuestra cada día que el balonmano da una afinidad especial; a Tiffany, porque me es imposible no quererla y sé que siempre estará ahí; a Balma, cuya transparencia me alegra cada día que la veo; a Paula, quien, al igual que a ella Barcelona, "em va robar el cor"; a Lluis, la bondad hecha persona; a Manuel, nuestro capitán planeta con ese humor que me vuelve loco; a Alba y a Ana, por su interés genuino, sus ganas de ayudar y por entenderme en mis momentos "fan-boy"; a Rafa, Eva y Jordi, por ser los sustentos silenciosos sin los que el lab se desmoronaría; a Dani Linares, del que no me puedo alegrar más que esa desconfianza científica de sus frutos al fin; a Jaime, por ser un referente de pensamiento crítico y trabajo duro; a mi tutor Josep Dalmau, por su diligencia y por siempre disponible ante cualquier cosa que necesitase; a mis "alemanas" Heike y Melanie, a las que me honra haber contribuido en hacerlas un poco más catalanas; a Lejla y a Carles, por su profesionalidad y serenidad en el día a día; a Alexis por sus recomendaciones y por ese puntillo malagueño que alegra a cualquiera; y a los más fugaces, Deborah, León, Lorena, Konrad, David, Adrià, Yerko… con los que no me importaría haber compartido muchos más momentos; y a todos los miembros de la comunidad Barccsyn con los que he tenido el privilegio de compartir tantos ratos: gracias Klaus, Maria Alemany, Cristina, Amanda, Vicky, Alex R., Alex H., Federico, Nico, Pasha, y a tantos otros que he conocido por el camino.

Finalmente, mi profundo agradecimiento para aquellas personas que, fuera del ámbito académico, han tenido un peso capital en mi vida. Esta tesis no se explica sin ti, Laura. Decirte que has sido la persona más importante de esta etapa es una obviedad tan grande que me da incluso vértigo al recordar los momentos que hemos compartido estos maravillosos años. Aunque a veces la vida carece de buenas explicaciones, te agradezco todo el amor que, tanto tú como tu familia, me habéis dado durante todo este tiempo. Gracias a los amigos "de siempre": mi retomado

Carles, quien me hace dudar de si mi afición por el ajedrez es por el mismo juego o porque me recurada a él, a Maria, por 10 años compartiendo pupitre, y a Cristina, por haber sabido aguantarme y entenderme de mil y una maneras distintas, siempre tendrás mi cariño. Gracias también a mis compañeros de máster Elena y Edu, con los que tanto me gusta compartir alegrías y penas, así como a mi querida "Rachel", ¡porque quién nos diría que unas tapas de boli y un rubio cobrizo nos llevarían tan lejos! Gracias también a mis maguitos de la SEI, a los más serios como mi querido maestro Alfredo, ¡del que tanto aprendo! o Xavi Soler, Amílkar, Eduard, Oriol, Jordi… ; y a los más gamberros, cuyas cenas me alegran la semana (David, Ernest, Manolo, Edu, Dani, Samu, Michelle, Juan, Sergio, Gabi, Tino…). De todos ellos, por supuesto, hay que destacar al otro "profe" (a Kike y Miguel ni medio gracias…) y mejor mago del mundo para mí, gracias por todo Pol. Y por supuesto gracias a los eternos "Tossencs": Rosa, Anna, Sandra, Delfi, Alba, Marione… por poder contar siempre con vosotros y por estar ahí cuando os necesito. En especial también gracias a Rosa por ser el puente hacia la persona más noble y pura que he conocido nunca. Tu nombre refleja lo que eres, Ester, y espero que estés siempre cerca en esta nueva etapa que ahora empieza. Por último, gracias también a mi otra familia, la familia del balonmano, a la que veo más que a la propia y que estos meses ha cogido todavía más importancia a mi vida. Entenderéis que destacar a alguno no sería de buen compañero y que todos, de maneras distintas, me habéis aportado algo por lo que merecía hacer siempre un esfuerzo más.

Si la tesis refleja a la persona, y las personas somos a su vez el espejo de las que elegimos que nos rodeen, entonces estoy tranquilo.

Gracias a todos.

1.Introduction

Working memory: first words

When I asked my mom to give me an example of "memory", she told me an event about her childhood. When I then asked her if she would consider as an example of memory when our rabbit "Yuri" always hided in the same place when any new guest came home, she doubted and decided to call it "instinct". When I finally asked her if she would consider as an example of "memory" the signature she wrote on a playing card for one of my magic tricks, she basically ignored me and told me not to bore her with stupid questions. This personal experience illustrates to what extent people tend to associate memory with a complex and exclusively human cognitive ability, while ignoring that a broad definition of memory would refer to the capacity for storing information; and animals, plants or even materials can do that.

And what about short-term memories? Afraid of my mom, I decided to ask my dad for an example of it, and he mentioned the last time he had to remember a telephone number and he kept repeating it in his mind until he could write it down… After interviewing my family, I remembered that when I started to be interested in Neuroscience, back in high school, understanding how we store information for just a few seconds attracted me more than understanding how we store information for years… and why? First, because, as opposed to brain, non-biological structures do not have different mechanisms for storing information at different timescales: a stone does not have one mechanism to hold color pigments for millennia and another to hold them for just a few seconds. And second, because it initially felt redundant: if we already have a mechanism to remember in the long-term, why do we need something else? is it just a failure of the main mechanism? The biological singularity and the apparent futility of this type of memory motivated a journey that started some years later with one term: "working memory".

Working memory (WM) is defined as the ability to maintain and manipulate information in the short-term. Chess is a great example of a continuous use of WM. During the 9th game for the world championship 2021, the defending champion Magnus Carlsen probably thought something like this: "if he moves the pawn to c5, then I can trap his bishop in b7 by moving my pawn to c6 and take that bishop with my rook three moves later". Magnus was running a kind of visual simulation and needed

WM to remember all the simulated moves. As usual, he excelled in using his WM, and found the simulated scenario that led him to win the game and, few games later, retained the title.

The term "working memory" was originally coined in 1960 by George Miller, Eugene Galanter and Karl Pribram. However, Alan Baddeley is the psychologist who defined it in its most common usage in the early 1970s (Baddeley & Hitch, 1974). He posited three components of WM: one responsible for storing verbal information (phonological loop), one responsible for storing visual information (visuospatial sketch pad), and one central component coordinating the other two (central executive). When Magnus Carlsen is remembering chess positions, he is using the visuospatial sketch pad; and when my father is remembering the telephone number, he uses the phonological loop. To coordinate them, the central executive comes in handy.

WM has shown to be very correlated with problem-solving abilities and intelligence. The amount of information that we can retain in the short term is called *WM capacity,* and several studies have correlated it to intelligence quotient (IQ) or general fluid intelligence (gF) (Alloway & Alloway, 2010; Gignac & Weiss, 2015; Kyllonen & Christal, 1990; Salthouse, 2014; Shipstead et al., 2016). However, these correlations are not straightforward, and many psychologists differentiate short-term memory and WM depending on the necessity to manipulate information (Klingberg, 2009). This ongoing debate expose one of the problems of defining cognitive processes from the psychological perspective: they usually rely on human interpretations and precise definitions instead of mechanistic observations of the brain function. In this thesis, I am not making a distinction of both terms and I am always referring to WM, as I studied it mechanistically through the development of computational models.

Computational models of WM? If explaining WM to my relatives is challenging sometimes, explaining that my thesis consisted in developing mathematical equations that reproduce how the brain stores memories, was epic. Furthermore, there was always the same follow up question: "what is it for?" and I must admit that I failed to transmit the relevance of the computational work until I watched a F1 race with my brother-in-law. When I was 12, I was a big fan of Fernando Alonso, and I spent several hours playing a F1 videogame with his official wheel and pedals. I used to

play the “championship mode”, which replicated the whole year competition, and I used the Friday’s practice sessions to get used to the circuit and make changes in the car. When my brother-in-law explained that the practice sessions were now much shorter and that some pilots just took a few laps, I was shocked. Apparently, drivers prepare the race in a simulator, getting used to the circuit and adjusting the parameters of the car in this virtual environment (MercedesBenz, 2020). By doing so, they avoid deteriorating the tires and do not expose the car to a crush. I suddenly realized this was a perfect analogy to computational modeling; F1 pilots can do that because they have extremely realistic simulators, and every change in the parameters of the car (aerodynamics, suspension, breaks, etc..) translates into the real world very accurately. That way, they can try more car configurations in less time and chose the ones that work better in that circuit. Computational modeling of WM is basically the same: instead of modeling a car in a circuit, we model the brain network responsible for storing information in the short-term. Although neuroscience is still far from making a model of the whole brain where the effect of a drug could be tested or the efficacy of a learning protocol evaluated, my work here goes in the same direction: describing the mechanisms of WM to develop models that contribute to a better understanding of the brain.

Working memory limitations

WM definition is circumscribed by its limitations. First, WM can just maintain a *limited* amount of information and, second, it maintains this information for a *limited* amount of time.

In the previous section I have already introduced the limitation in the amount of information: WM capacity. In his 1956 article “The magical number seven, plus or minus two” (G. A. Miller, 1956), George Miller hypothesized a fixed capacity for the human ability to receive information, and that this limit was around seven items. Posterior experiments using WM change-detection tasks (instead of memorizing multiple items, detecting changes in their features when seeing them again), found a much lower capacity of between 3 and 4 items (Cowan, 2001; Luck & Vogel, 1997).

Wilken & Ma (2004) observed that precision to remembered items decayed with WM load (number of items to remember) and proposed that WM consisted of a pool of resources that can be allocated flexibly to provide either a small number of high-resolution representations or a large number of low-resolution representations (Frick, 1988; Wilken & Ma, 2004) – the "resource model"-. An alternative hypothesis proposed that WM stores a limited set of discrete, fixed-resolution representations -the "slot model". As an analogy, consider "resource model" as three cups (items to remember) and a bottle of juice (WM resources), so most of the juice could be poured most into a single cup, leaving only a few drops for the other two cups. In the "slot model", WM resources are like a set of prepackaged juice boxes, so a cup could never receive just "a few drops" (W. Zhang & Luck, 2008). In 2008, two studies simultaneously published convincing evidence towards each model (Bays & Husain, 2008; W. Zhang & Luck, 2008). While Zhang & Luck (2008) found that the "slot model" fitted behavioral data more accurately when cuing one specific stimulus (Figure 1A-B), Bays & Husain (2008) commented that the previous

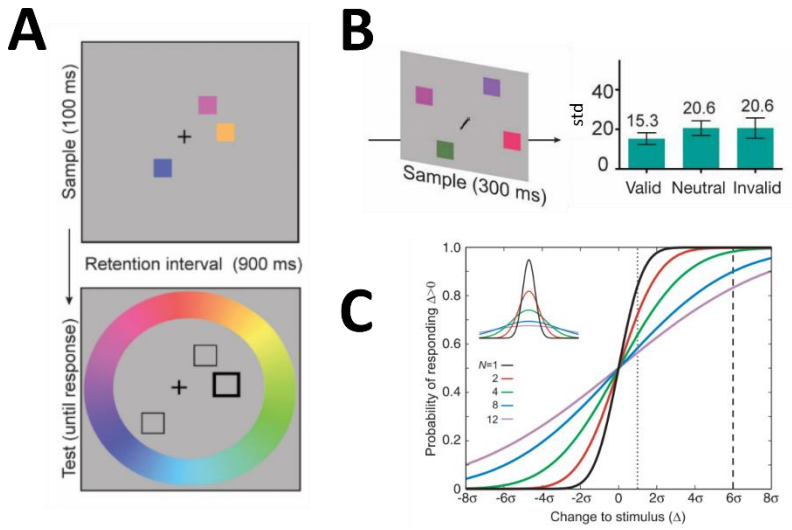

Figure 1 WM capacity

A) Schematic view of a load 3 trial (3 stimuli to remember) in the WM task of Zhang & Luck (2008), where participants had to remember the color at each location. **B)** When previously cuing one of the stimuli (70% of the time the subject will be asked to report the cued stimulus), no dramatic change in the standard deviation (std) is observed between the cued stimulus (item) and the rest (invalid and neutral), which is not consistent with the "resource model" but it is with the "slot model". **C)** From Bays & Husain (2008). Response probability as a function of the size of the change to the stimulus for different numbers of items. The curves become flatter with increasing number of items, corresponding to changes in the Gaussian distributions of error.

experiment needed to control for eye movement, which was critical to observe that, indeed, the "resource model" explained better the precision for very high loads (Figure 1C).

The other source of limitation regards to the fidelity of WM with time. The interval of time from the last time the WM content is available to the senses and the time to make use of it to guide behavior is called *delay*. The first experiments measuring how WM precision decayed with increasing delay were done by Friederich Hegelmaier during the XIXth century (Laming & Laming, 1992). And from there, many studies have replicated

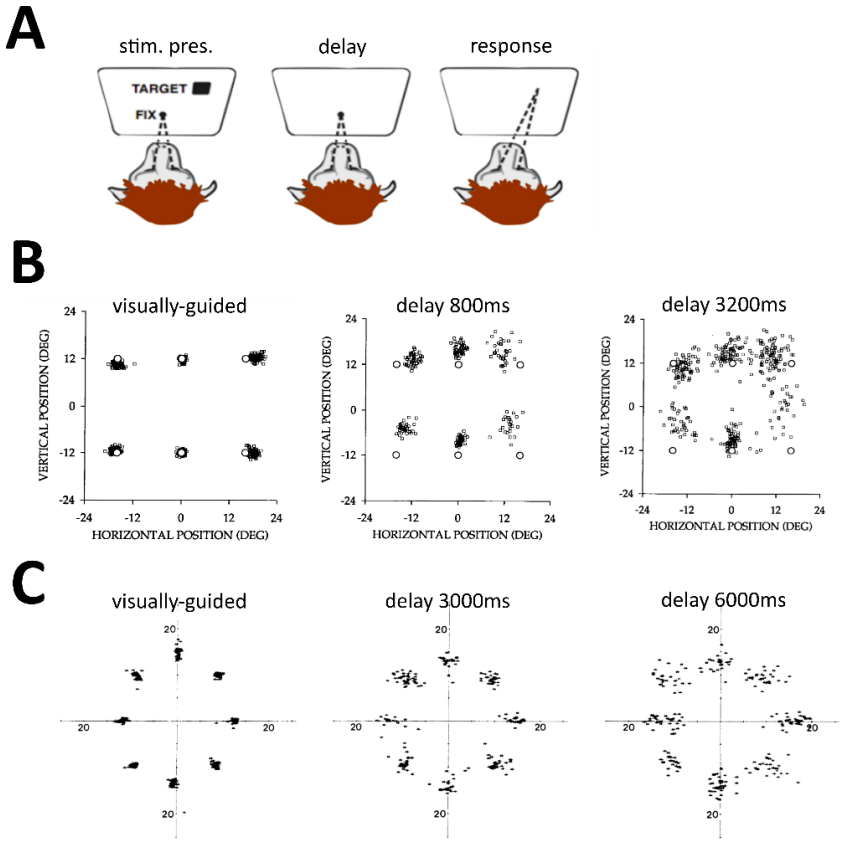

Figure 2 ODR task and delay effects

A) Schematic view of the oculomotor delayed response task (ODR task) in a monkey experiment. The monkey must fixate, a stimulus appears in the periphery and the monkey must remember its location during the delay time. When the fixation point disappears, the monkey makes a saccade to the remembered location to receive reward (juice). B) Precision of the saccades with increasing delay length in White et al. (1994). End points of remembered saccades are less accurate and more scattered than those of control saccades (visually-guided: no delay, $p<0.01$). C) Precision of the saccades with increasing delay length in Funahashi et al. (1989).

similar effects with delay length (Barrouillet et al., 2012; McKeown & Mercer, 2012; Pertzov et al., 2013; Vergauwe et al., 2010). In this thesis, I will evaluate visuospatial WM (vsWM), and the most classical task to test it is the *oculomotor delayed-response* (ODR) *task*. In it, subjects must fixate on a central stimulus while a target appears in the peripheral space. Subjects must memorize the location of the target, and after a varied delay period, the fixation cue disappears. The disappearance of the fixation point cues the participant to make an eye movement to the location in which the target had appeared. The simplicity of this task made it suitable for experiments both in humans and animals (Funahashi et al., 1989, 1991, 1993; Goldman-Rakic et al., 1990; Ploner et al., 1998; White et al., 1994). Figure 2A shows a schematic view of the ODR task and Figure 2B-C the decrease in precision with increasing delay lengths in White et al. (1994) and Funahashi et al. (1989), two experiments made in monkeys.

WM is related to attentional processes (De Fockert et al., 2001; Fougnie, 2008; Konstantinou et al., 2014; Spinks et al., 2004). Spinks et al. (2004) showed that only under high WM demands (complicated mathematical operations), activation in early sensory areas was impaired, demonstrating modulatory top-down effects. On the other hand, bottom-up effects of saliency influencing information processing are well described (Itti & Koch, 2001; Posner, 1980; Wolf & Lappe, 2021). The relation between attention mechanisms and WM is clear when introducing distractors in WM tasks. A good example of this balance is the "cockatil party effect". When you are standing in the middle of a group of chatting people, you concentrate your WM on the person to whom you are talking, and attention filters out all

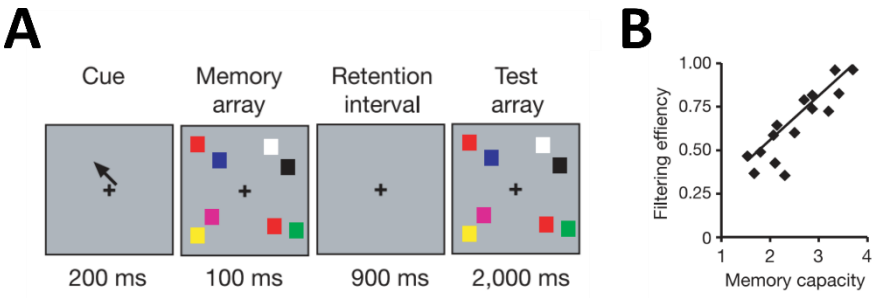

Figure 3 Distractor filtering and WM capacity
Modified figure from Vogel et al. (2005). **A)** Participants were asked to remember the colors in the top left hemisphere. They were tested 1s later with a test array that was either identical (example) or different. **B)** Correlation between the efficiency in filtering out distractors and the WM capacity: individuals from the high WM capacity group filtered out distractors more efficiently.

the conversations going on around you. However, if someone behind you mentions your name, you cannot avoid being distracted. Illustrating the tight relation between attention and WM, people differ in how they perform in the cocktail party situation depending on their WM capacity: those with the lowest WM capacity are the most easily distracted (Conway et al., 2001). Figure 3 shows how this results have been replicated in vsWM tasks (Vogel et al., 2005), showing that in individuals where WM fails due to low WM capacity, distractors take over easily (Kane et al., 2017; Klingberg, 2009).

In this thesis, I will explore how the WM content is affected by capacity, delay, and distracting information. To do so, I will simultaneously manipulate the spatial and temporal domains with relevant (target) and irrelevant (distractor) stimulus. Regarding the spatial domain manipulations, I will study how the distance to the fixation point (eccentricity) affects WM content. Previous studies showed a compression of the visual space along the radial dimension (*foveal bias*) during steady eye fixation (Cai et al., 1997; Honda, 1993, 1995; Kerzel, 2002; Mateeff & Gourevich, 1983; Mewhort & Campbell, 1978; Mitrani & Dimitrov, 1982; Osaka, 1977; Ross et al., 1997; Townsend, 1973). This effect consists of systematic mislocalizations towards the fixation point in estimating the position of briefly presented targets. This effect gets magnified with eccentricity, so the farther the target is presented from fixation, the higher the mislocalization is (Müsseler et al., 1999). Sheth & Shimojo (2001) deeply explored this effect and discovered that it was not restricted to perception. They observed that mislocalizations towards fixation increased with delay length (Sheth & Shimojo, 2001).

One of the most consistent findings in WM capacity limitations is that high featural overlap between memories generates more memory impairment than low featural overlap. This has sometimes been defined as “similarity” or “congruency” effects (Lorenc et al., 2021), and it is observed, for example, when WM for faces is more impaired by other face distractors than by scene distractors (Yoon et al., 2006). The spatial distance separating visual stimulus is also a domain of similarity, but previous studies showed some complexity on it. Manipulating the distance between presented items (target-target distance and target-distractor distance) showed the remembered locations could be biased towards each other (*Memory averaging*, Figure 4) (Barbosa et al., 2020; Brady &

Alvarez, 2011; Elmore et al., 2011; Herwig et al., 2010; Hubbard, 1995, 1998; Hubbard & Ruppel, 2000; Johnson et al., 2009; MacOveanu et al., 2007; Stein et al., 2020; Van Der Stigchel et al., 2007) in different spaces such as the color-space (Barbosa & Compte, 2020; Teng & Kravitz, 2019) or the orientation-space (Chunharas et al., 2019; Lorenc et al., 2018; Rademaker et al., 2015, 2019) but also repelled (Bae & Luck, 2017; Fritsche et al., 2017; Kerzel, 2002). In a very important study for this thesis, Almeida et al. (2015) described both attractive and repulsive biases (Figure 4C) between simultaneously presented targets depending on the distance: attraction for close-by locations and repulsion for distant ones (Almeida et al., 2015).

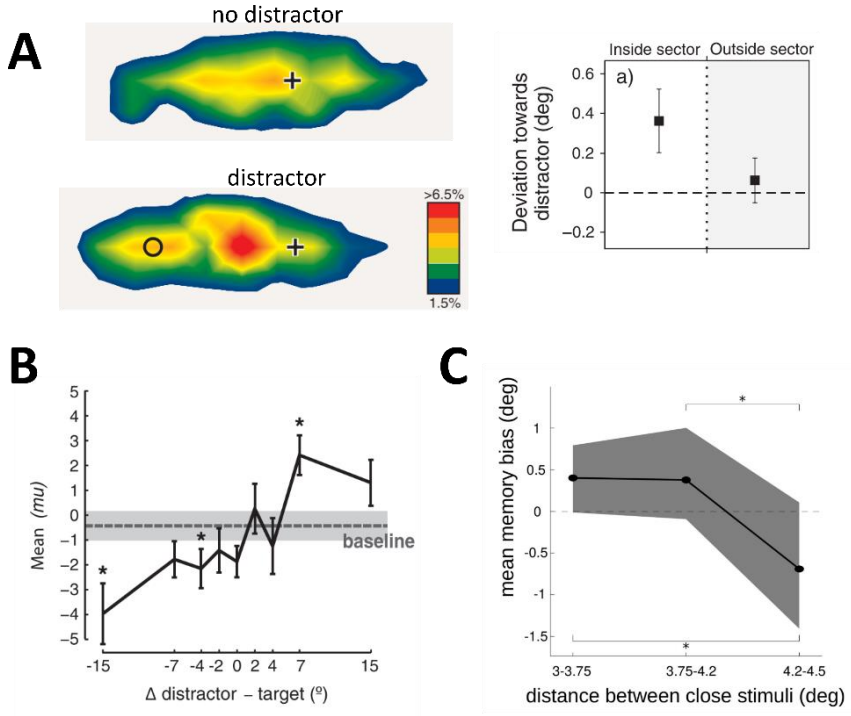

Figure 4 Distance effect in memory reports

A) Modified figure from Herwig et al. (2010). On the left, probability maps of the saccades' landing positions for targets as a function of distractor condition. On the right, the deviation towards the distractor is shown, just for distractors located in the same sector (±20° around the target). **B)** Figure from Rademaker et al. (2015), where they showed attraction towards the orientation of a distracting grating presented during the delay period. **C)** Figure from Almeida et al. (2015), showing attraction between simultaneously presented target during the delay period when they were located close (<4.2° of visual angle) and repulsion when this distance increased.

Regarding the temporal domain, in this thesis I manipulate the delay time to compare perceptually-related errors at vanishing delays to WM-related errors. Furthermore, I vary the time between the stimulus presentations to define the dynamics of distractor interference in WM. When the time between the target and the distractor was studied parametrically, distractors presented short time after the target presentation -short *target-distractor onset asynchrony* (TDOA)- had more distracting effect than those with long TDOA (Suzuki & Gottlieb, 2013). Literature regarding the temporal manipulation of the distractors is rare (Jolicœur & Dell'Acqua, 1998; McNab & Dolan, 2014; Pasternak & Zaksas, 2003; Suzuki & Gottlieb, 2013; Van Ede et al., 2018; Vogel et al., 2006), but still consistent regarding short TDOAs being more distracting. To my knowledge, Murray et al. (2017) is the only one proposing a mechanism for this specific effect mediated by frontal and parietal regions that explains this effect, although previous studies supporting an attentional "gate mechanism" mediated by the basal ganglia could also explain it (Frank et al., 2001; McNab & Klingberg, 2008). Combining the temporal with the spatial manipulation provides comprehensive constraints for a computational understanding of the circuit mechanisms of WM maintenance and robustness to distraction.

Working memory in the brain

Studying the effects of brain lesions on behavior is one of the most established and influential methods in neuroscience, and constituted the foundation of cognitive neuroscience in the mid to late 20th century (Vaidya et al., 2019). A vast number of independent studies, both in humans and monkeys, show that lesions in prefrontal cortex (PFC) impair WM (Chao & Knight, 1998; Fuster, 1988; Goldman-Rakic, 1987; Jacobsen, 1936; Lara & Wallis, 2015; Milner, 1963; Karl H. Pribram et al., 1952; Voytek & Knight, 2010; Warren et al., 1957). In 1971, two independent groups published the results of monkeys trained to perform a WM task while recording the activity of PFC neurons using micro-electrodes previously introduced into their brains (Fuster & Alexander, 1971; Kubota & Niki, 1971). Both studies found that during the delay period some neurons fire constantly at an elevated firing rate (persistent activity, PA), providing the first evidence of a neural correlate of WM. The

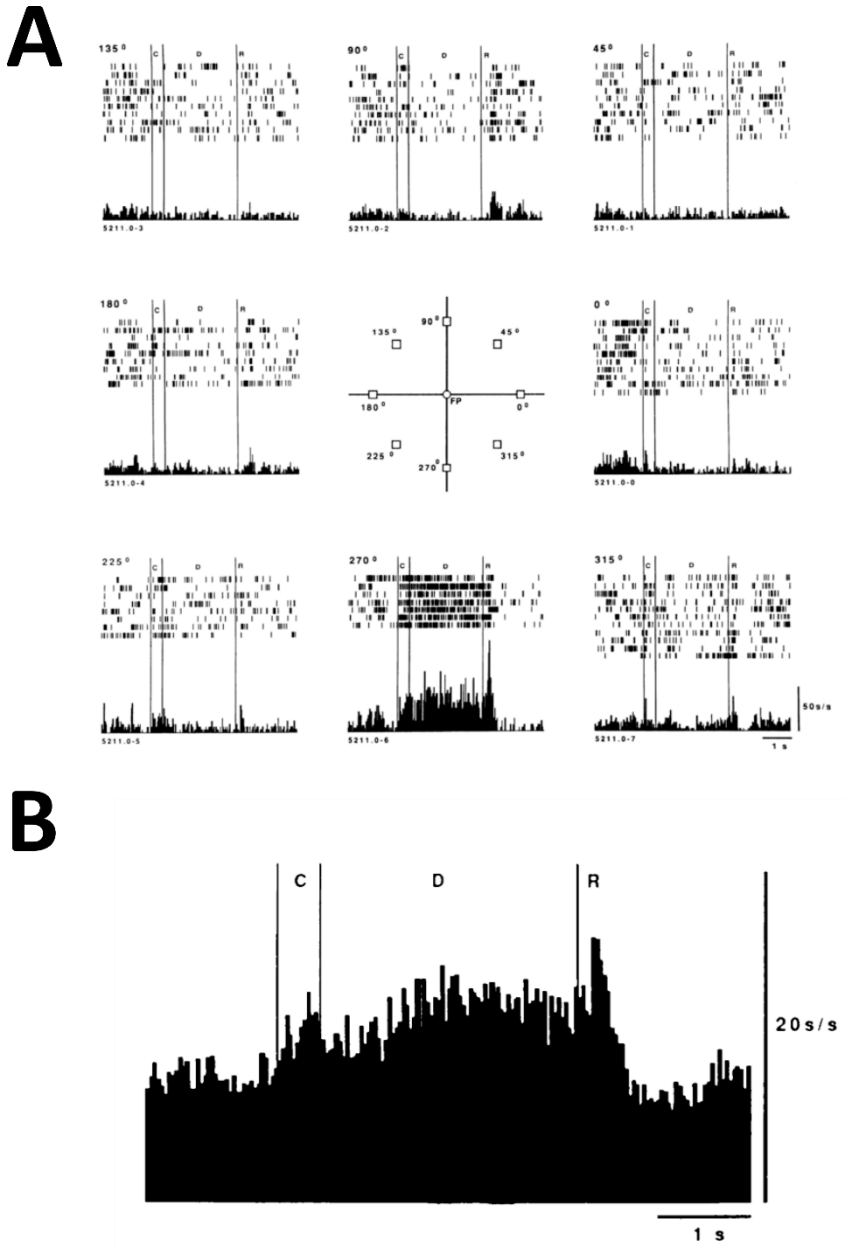

Figure 5 Persistent activity and selectivity to spatial locations

Figure from Funahashi et al. (1989). **A)** Neuron recorded in PFC during a vsWM task. Each panels shows the neural activity of the same neuron when a stimulus is presented in one of the 8 possible configurations (0°-315°). The neuron presented PA during the delay period just when remembering the 270° location. In each panel, the first two vertical lines indicate the presentation of the stimulus and the las one, the response time. The histograms sum the neural activity in the different trials (raster plots on top of each). **B)** Histogram of the sum neural activity at the preferred location of 46 neurons with excitatory directional delay period activity aligned at the cue presentation.

interpretations of both groups, however, were slightly different. While Fuster & Alexander (1971) identified the pattern of activity as the neural correlate of the memory, Kubota & Niki (1971) thought that the activity was related to motor plan (future monkey's movement). Twenty years later, the laboratory Patricia S. Goldman-Rakic addressed this controversy by changing the motor plan on a trial-by-trial basis with saccades towards the remembered location and saccades in the opposite direction and still found PA in the same prefrontal neurons (Funahashi et al., 1993). This insight was possible because of the advancement that had supposed, a few years before, the introduction of the finely controlled ODR task for the investigation of the neural bases of WM (Funahashi et al., 1989). In this work, they trained two monkeys to perform an ODR task, with the visual stimulus placed at one of 8 possible different spatial locations. This time, apart from observing PA, they found selectivity to the memorized

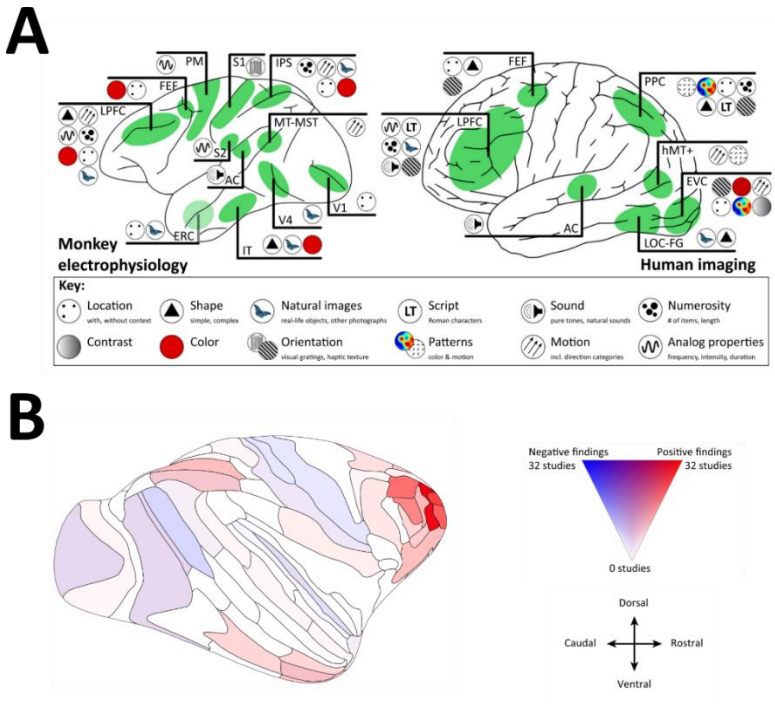

Figure 6 Reviews of WM-related PA

A) From Christophel et al. (2017). Summary of all areas showing delay activity depending on the WM content. The left plot shows the findings in the monkey brain and, the right one, the findings in the human brain. B) From Leavitt et al. (2017). Meta-analysis confronting negative (blue) and positive (red) findings of PA in the monkey brain. Results reveal that robust PA is mainly found in downstream areas such as parietal and frontal cortex.

locations (Figure 5). The finding of tuned PA and the insights from lesions study establish PFC as a candidate area for WM maintenance. However, WM-related PA has also been observed in many other brain regions (Chelazzi et al., 2001; Gnadt & Andersen, 1988; Pesaran et al., 2002; Supèr et al., 2001). In 2017, two extensive reviews summarized these findings (Christophel et al., 2017; Leavitt et al., 2017). Christophel et al. (2017) nicely showed, qualitatively, how different WM content has been observed in different brain regions while Leavitt et al. (2017) quantitatively reviewed more than 90 studies studying PA associated to WM, observing PA is a much more stable correlate of WM in frontal areas compared to sensory areas.

Evidence from *functional magnetic resonance imaging* (fMRI) and the application of decoders to neuroimaging data (Figure 7), questioned the direct electrophysiological recordings of PFC, as they reveal that visual perception and WM maintenance of a visual object elicit qualitatively similar patterns of neural activity in early sensory regions (Albers et al., 2013; Christophel et al., 2012; Gayet et al., 2017, 2018; S. A. Harrison & Tong, 2009; Serences et al., 2009; Serences, 2016). Using refined neuroimaging techniques such as the inverted encoding models (IEM), which reconstruct WM content from patterns of Blood Oxygenation Level Dependent (BOLD) activity, highlighted the importance of sensory areas in vsWM (Figure 7B- D), correlating reconstruction strength with behaviour (Emrich et al., 2013; Ester et al., 2013; Hallenbeck et al., 2021; Lorenc et al., 2018; Rademaker et al., 2019; Sprague et al., 2014, 2016). These results motivated a new theory for WM maintenance: the *Sensory Recruitment Theory* (SRT), which states that the same areas responsible for the encoding are also responsible for the maintenance in vsWM. However, not all fMRI data is consistent with this theory. Bettencourt & Xu (2016) showed that although information about the target can be extracted from visual areas (V1-V4) during the delay period, it disappear from visual cortex when distractors are introduced during the delay, with information remaining in the parietal cortex (Bettencourt & Xu, 2016). Some critiques regarding the efficiency of the distractors were made (Gayet et al., 2018; Scimeca et al., 2018) and other experiments indeed found decoding in visual cortex under distraction (Hallenbeck et al., 2021; Lorenc et al., 2018; Rademaker et al., 2019), so the role of these regions during WM maintenance under distraction is still under debate. In this thesis, I

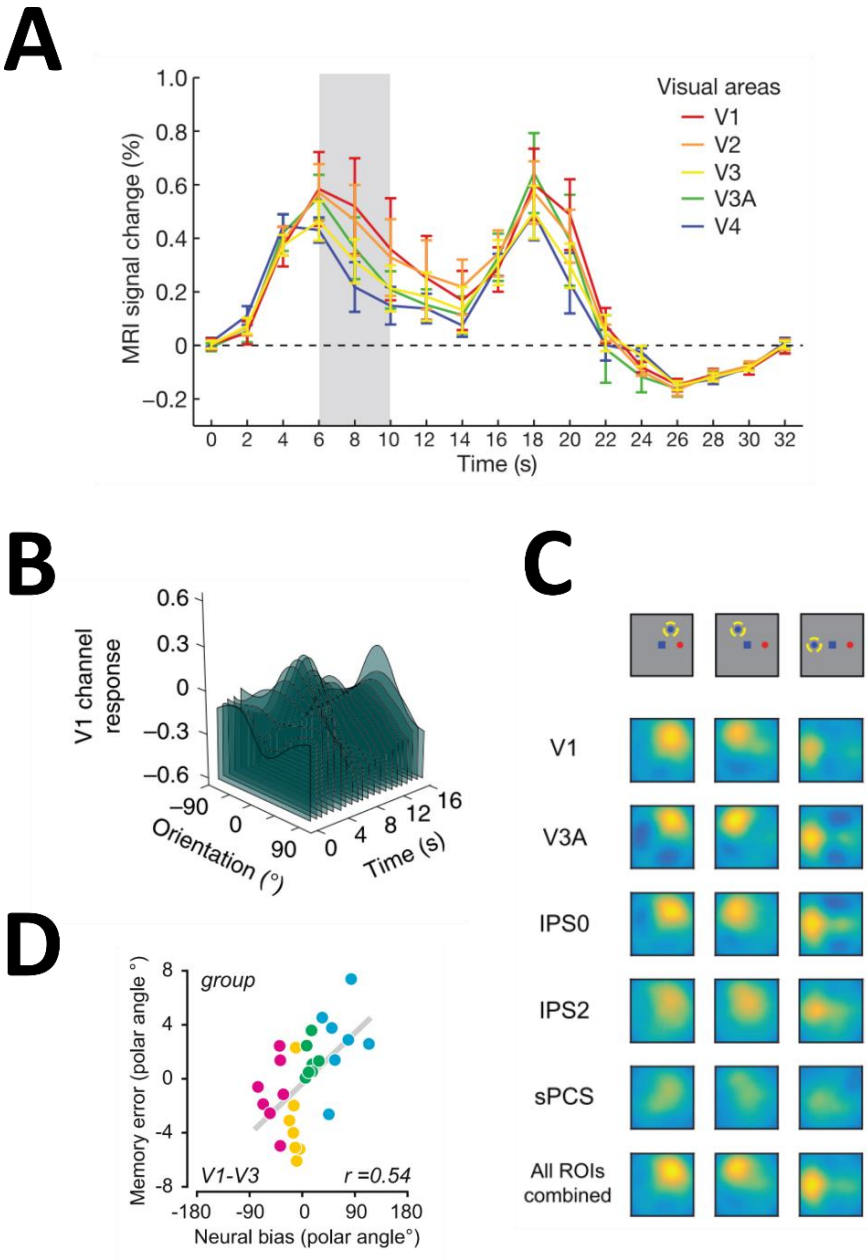

Figure 7 fMRI data in support of the Sensory Recruitment Theory

A) Figure from Harrison & Tong (2009), showing the time course of mean BOLD signal (n=6) in early sensory areas (V1–V4) during a WM task. The shaded area indicates an interval inside the delay period, which started at 2s and lasted 11s. **B)** Figure from Rademaker et al. (2019), showing the time course of the reconstruction of an orientation in V1 using an IEM. **C)** Figure from Sprague et al. (2016), showing the reconstruction during the delay period of two simultaneous memories in different regions. **D)** Figure from Hallenbeck et al. (2021) showing a positive correlation between the reconstructed error and the actual error in early sensory areas.

evaluate the SRT, first by reconstructing target-related and distractor-related WM content from fMRI signals in different regions and distracting conditions and, second, by testing a fundamental point of the theory, which is the predicted interference based on the shared neural circuit (Fischer & Whitney, 2014; W. J. Harrison & Bays, 2018; Teng & Kravitz, 2019): as a neural circuit is shared, perceptual responses based on immediate reaction upon incoming information should present the same topographically-based biases as responses in delayed paradigms that require WM.

Computational models of working memory

Developing computational models of cognitive processes is one of the ultimate goals of neuroscience, as they contribute to a mechanistic explanation of the human brain. The previously presented literature supporting that PA in frontal regions is a neural correlate of WM (Funahashi et al., 1989, 1993; Fuster & Alexander, 1971; Goldman-Rakic, 1995; Kubota & Niki, 1971) supported the postulations of Lorente de Nó and Donald Hebb about a central neural mechanism for the maintenance of information during the delay period mediated by reverberatory activity in synaptic feedback loops (Hebb, 1949; Seung, 2000). This framework motivated models to explain WM maintenance through reverberatory activity within a local recurrent neural network that displays bistability between a resting state -no memory- and a structured activity state -maintenance- (Amari, 1977; Amit, 1995; Amit & Brunel, 1997; Wilson & Cowan, 1973).

The bump attractor model (Compte et al., 2000) has been very successful in replicating both behavioral and electrophysiological data (Almeida et al., 2015; Wimmer et al., 2014). Simulations of the bump attractor model (Figure 11, *Methods-Computational modeling*) reveal that activity in the network behaves as a continuous attractor. The activity peak, or “bump”, formed by the firing rate of the population of neurons organized according to neuronal memory selectivity survives the disappearance of the stimulus, but the location it represents diffuses slowly in time away from that of the original location of the stimulus (Burak & Fieted, 2012; Compte et al., 2000). The model has been able to replicate many of the previously

presented limitations of WM. When the WM load increases, a localized persistent activity bump may either fade out or merge with another nearby bump (Z. Wei et al., 2012). Interference effects have also been through the connectivity of excitatory and inhibitory neurons (Almeida et al., 2015; Nassar et al., 2018), and the loss of precision with delay is explained by the diffusion of the bump during the delay period. In this line, Wimmer et al. (2014) showed that the reported saccades of monkeys in a vsWM task deviate consistently in the direction of the bias decoded from neuronal activity recorded from single neurons at the end of the delay period. The bump attractor has a very defined topographical configuration: “a ring”, where each neuron memory field covers different angular locations in the visual field at a constant distance from the fovea. Therefore, the previously reported effects have never been studied in computational model for different distances from the fovea. In this thesis, I will extend the topographical interpretation of the circuit in the angular dimension, by exploring angular effects in interference and diffusion at different eccentricities; but also in the radial dimension, by developing a new radial model where memories drift not in the angular dimension but the radial dimension, aiming to mechanistically explain, for the first time, the documented compression of the visual field in WM (Sheth & Shimojo, 2001).

Alternative mechanistic explanations for memory maintenance have also been proposed. Neuroscientists defending the SRT, for example, need alternative explanations for memory maintenance, as PA in visual regions has been seldomly reported, and when observed it is usually very weak (Leavitt et al., 2017; Supèr et al., 2001). Some models of WM rely on synaptic processes -*activity silent* (no PA)- such as changes in presynaptic calcium levels, or fast Hebbian synaptic plasticity (Mongillo et al., 2008; Sandberg et al., 2003; Sugase-Miyamoto et al., 2008). Experiments using retro-cues (a cue during the maintenance period that indicates which of multiple items is the relevant one) provided some indirect evidence for activity silent mechanisms of WM for memories that are not in the focus of attention. Two independent studies (Rose et al., 2016; Wolff et al., 2017) showed that WM decoding from brain activity is much stronger when the remembered stimulus is in the focus of attention than when it is not. When a remembered stimulus leaves the focus of attention, decoding drops but, critically, it recovers when the stimulus comes back into the focus of attention. More powerful analyses suggest that unattended

memories are characterized by weaker, not absent decoding (Barbosa et al., 2021; Christophel et al., 2018), which is consistent with analysis of spiking activity showing significant PA or decoding for unattended stimuli or distractors (Jacob & Nieder, 2014; E. Miller et al., 1996; Panichello & Buschman, 2021; Rainer et al., 1998; Spaak et al., 2017; Watanabe & Funahashi, 2014). This weakens the value of this data as evidence for activity-silent WM, but interaction between PA and synaptic storage mechanisms could still be at the basis of the difference in neural activity between attended and unattended memories. Indeed, computational models of PA have been shown to benefit of other mechanisms of maintenance or of the regulatory action of other areas to explain behavior.

One example of this, is the interference occurring between previous responses and the forthcoming stimuli (Underwood, 1957), with similar attractive and repulsive effects as the previously presented "similarity effects" (Bliss & D'Esposito, 2017; Fischer & Whitney, 2014; Fritsche et al., 2017; Papadimitriou et al., 2015). To successfully replicate this behavioral effect, the bump attractor model needs to incorporate slower cellular or synaptic mechanisms, such as cannabinoid-mediated disinhibition (Carter & Wang, 2007) or short-term synaptic plasticity (STP) (Barbosa et al., 2020; Kilpatrick, 2018; Stein et al., 2020). In this line, previous studies explored the combination of STP and PA mechanisms, showing that while short-term facilitation (STF) decreased diffusion with delay and biases towards distractors, short-term depression (STD) increased them (Hansel & Mato, 2013; Itskov et al., 2011; Seeholzer et al., 2019). Another example is the frontoparietal circuit model proposed by Murray, Jaramillo et al. (2017), where they explain the TDOA effects of distractors (early distractors interfering more than late distractors through TDOA) through the mechanisms of interaction with other brain areas (Murray, Jaramillo, et al., 2017). In this thesis, I explore and develop new versions of the bump attractor model to explain behavior under different scenarios: with and without distraction, single and multiple stimuli, different eccentricities, varying delay length, interleaved-trial vs block design and under electrical stimulation. By doing so, I will show the flexibility of the bump attractor model and its biological fitness. Furthermore, I collect behavioral, electrophysiological and neuroimaging evidence that suggest that frontal areas and not early sensory areas are being responsible for the final memory readout through attractor dynamics.

2.Goals

The main goal of this thesis is to study the circuit for WM maintenance from a mechanistic perspective. To do so, I combine behavioral experiments with neuroimaging techniques and neuronal recordings under the framework of the bump attractor model. The work performed during this thesis is contained in two main chapters, one focusing on describing the topography of the WM circuit, and the other focusing on how distractors interfere with the WM content mechanistically.

In the Chapter *Topography of the working memory circuit*, I test assumptions of the Sensory Recruitment Theory (SRT) by focusing on an analysis of the topography of WM. My goals are: **(1) to determine if topographical relationships are maintained through encoding and maintenance periods of WM, as expected if they share the same neural circuit**. This will be done by changing the eccentricity, the angular separation between memoranda and the delay length in a parametric way; and **(2) to extend the bump attractor model to incorporate topographical features of WM on the radial dimension.** This will be accomplished by using computational modeling to reproduce and interpret the behavioral results obtained in Goal 1. Thus, this chapter test the SRT, it will provide evidence towards memory maintenance mediated by attractor dynamics, and it will propose a biologically plausible mechanistic explanation for memory effects both in the angular and the radial dimension.

In the Chapter *Distractor filtering in the working memory circuit*, my goals are: **(3) to evaluate the effect of distractors in the similarity and temporal domains in behavior and neuroimaging.** These results will be interpreted in the context of the SRT and previous literature on WM decoding using IEMs; **(4) to propose mechanisms for distractor filtering in WM using a computational network model framework;** and **(5) to test predictions arising from computational models in neural datasets**. To this end, I analyze different datasets following predictions from the models.

3.Methods

Paradigms and analysis

In this thesis, I run two different psychophysics paradigms to test vsWM. The first one was designed to study the topography of the circuit and its delay-dependence (Chapter 1: *Topography of the working memory circuit*) while the second was designed to study distractor filtering in vsWM (Chapter 2: *Distractor filtering in the working memory circuit*). Data acquisition, data processing and data visualization and statistical analysis were done using *python2 and python3* open-source libraries and custom-made code.

Paradigm Topography of vsWM

I developed a vsWM task in which 18 fixating subjects (11 female) had to remember the spatial location of a set of stimuli (colored dots) and report, after a delay period, the location of one of them (Figure 35, in *Results*). Each trial consisted of fixation, stimulus presentation, delay, and response periods.

The stimuli consisted of different colored dots with a radius of 0.4cm. The colors were chosen randomly in each trial from a palette of 5 colors: red, blue, green, white, and gold. Stimulus location was specified based on polar coordinates (radius, angle) with fixation as origin. The stimuli were displayed on a computer screen (38,61cm x 28,96cm), on a grey background for 500ms and they could be placed at one of the three different radii from fixation (radius1: 7.78cm, radius2: 10.70cm and radius3: 13.68cm). Depending on the number of dots to be remembered in each trial, I distinguish between two types of trials: multi-item trials (two dots presented simultaneously) and single-item trials (one dot presented alone). From a total amount of 8186 trials, 3396 were single-item trials and 4790 were multi-item trials. In both types of trials, subjects just reported the location of one of them. The location to report was cued at the end of the delay period, when the fixation point changed color to match one of the presented stimuli.

In multi-item trials, stimuli could be separated in the angular dimension (same radius, different angle: 2403 trials) or in the radial dimension (different radius, same angle: 2387 trials, not used for the thesis). When stimuli were separated in the angular dimension, I used 2 radii (radius 1: 7.78cm and radius 3: 13.68cm) and 3 angular distances (12°, 16° and 20°).

In half of the trials, subjects had to report the clockwise (cw) stimulus and in the other half, the counterclockwise (ccw) stimulus. When stimuli were separated in the radial dimension, I used the same angular distance, but different radii (radius1, radius2 and radius3). In all types of trials, I had a delay 0 condition, where the participants were asked to respond just after the presentation period finished, and a delay 3 condition, where just the fixation point was visible for 3 seconds before the response time (they did not see the stimulus during the response time in either case). Trials were randomly interleaved, so participants could not predict either the locations of the stimuli nor the delay period duration.

Participants completed the task in 3 different days. To avoid fatigue, participants completed 3 sessions of 58 trials per day, with breaks of ~10 min in between. Each round lasted ~15min. Before the experiment, the experimenter explained the task and the subjects could practice in pilot trials until they understood it. During the experiment, the participant's head was supported using a chinrest situated at ~47cm from the screen. They were seated in front of the screen, and they were asked to fixate the central black square during the fixation period, stimulus presentation and delay period. Participants were asked to break fixation during the response period. An eye tracker (pupillabs®, pupil w120 e200) was used to control for fixation. If participants broke fixation, the trial was invalidated. To avoid a massive loss of trials, invalidated trials were presented again later in the same session. If a trial was invalidated twice, it was removed. Participants used a pressure-sensitive tablet and a pen to respond. The movement of the pen was reproduced in the visual display as a cursor. To start a trial, participants had to set the pen in the visual display. A black square appeared on the visual display when they did it. They had to drag the black square to the fixation point and fixate on it. During stimulus presentation and delay period, participants were asked not to move the pen as well as not to break fixation. In the response period, the black square colored with the color of one of the presented stimuli. Participants reported the position of the asked stimulus by dragging the colored square to the remembered position and, once there, releasing the pen from the tablet. To limit possible pattern encoding strategies, I presented stimuli in the 60° around the diagonal of each quadrat, so that the geometrical symmetries or cardinal directions were avoided (Zelinski, 2016).

Paradigm Distractor filtering

I developed a vsWM task in which 27 fixating subjects (19 female) had to remember the spatial location of a set of 3 stimuli (targets) and ignore another set of 3 stimuli (distractors) presented separately. Distractors could be presented prospectively (order 1) or retrospectively (order 2) to the targets and the time between the targets and the distractors -target-distractor onset asynchrony (TDOA)- was also manipulated. The task had a short-TDOA condition of 200ms and a long-TDOA condition of 7000ms (Figure 43, in *Results*). Order and TDOA manipulations were combined and randomly interleaved, giving four different temporal conditions: *order 1 – short TDOA*, *order 1 – long TDOA*, *order 2 – short TDOA* and *order 2 – long TDOA.*

The task consisted in a sequence of trials, each one starting when the subjects moved the mouse on top of the fixation point. The fixation point was placed in the middle of the screen (diagonal=23.7cm, 1920x1080px), on a grey background. It consisted in an empty black square subdivided in four (side length of each sub-square = 0.3cm). Subjects were instructed that each sub-square represented the corresponding quadrant (top-right: first quadrant (0°-90°), top-left: second quadrant (90°-180°), bottom-left: third quadrant (180°-270°), bottom-right: fourth quadrant (270°-360°)). Once the subject fixated, a digit (order cue) indicating the set of stimuli to be remembered appeared. A 1 indicated subjects to remember the location of the first set of 3 stimuli (targets) and ignore the second set (distractors) while a 2 indicated subjects to ignore the first set of stimuli (distractors) and remember the location of the second set (targets). The order cue was presented for 500ms. If subjects missed or forgot the cue during the task, they were instructed not to move the cursor during the response period. Out of 6874 trials, this occurred in 112 trials (1.63%), which were excluded from the analysis.

All the stimuli (targets and distractors) consisted of black dots with a radius of 0.5cm. The stimuli were displayed on a computer screen on a grey background for 350ms. All the stimuli were presented on top of a black circle (r=8cm). All three targets were presented in different quadrants and they were never presented close to the cardinal axes (20°) to prevent from any perceptual confusion of the quadrant presentation and to avoid strategies based on references (Zelinski, 2016). Distractors were also presented in three different quadrants with equal restrictions to cardinal axes. The distance between targets and distractor were manipulated the following way: one distractor was in the same quadrant of one target,

separated by 10°-20°. Another distractor was in the same quadrant of another target, separated 20°-30°. The remaining distractor was in the quadrant without target. That way, two quadrants had both a target and a distractor and the remaining two quadrants had one target and one distractor alone, respectively. Besides controlling for the distance, I also controlled for the cw and ccw disposition of the distractors. For each target-distractor distance, the distractor was placed cw to the target in half of the trials and ccw in the other half. In each trial, the cw and ccw disposition of the distractors were randomly assigned (not all three distractors were cw or ccw).

The maintenance period of the vsWM task was 12s in all the conditions. In order 1 conditions, distractors were presented during the 12s of the delay while in the order 2 conditions, distractors were presented before the 12s of delay start. At the end of the delay period, one of the squares of the fixation point turned yellow, instructing the subject to report the location of the target located in that quadrant. Simultaneously, a yellow bar appeared on top of one of the closest cardinal axes (vertical or horizontal, randomly chosen in each trial). By using the scroll of the mouse, participants had to adjust the position of the bar and make a left click to confirm. Participants had a maximum time of 10s to respond.

21 participants (16 female) completed the task in the laboratory facility for psychophysical experiments and 6 participants (3 female) completed the task inside the scanner. Although the task was the same, the target separated 20°-30° to a distractor was never the one cued for response in the scanner setting. Participants doing the task in the laboratory facility made a total of 7 runs of 10-15min each, divided in two different days to avoid fatigue. Between runs, participants could make a break of 5-10min. Before the experiment, the experimenter explained the task and the subjects could practice in pilot trials until they comprehended it. During the experiment, the participant's head was supported using a chinrest situated at ~60cm from the screen. Participants were asked not to break fixation during the trial period. If they needed to rest between trials, they could move the cursor out of the fixation point and break fixation. Whenever they wanted to continue with the task, they were instructed to fixate their eyes in the fixation point and moved the cursor to it. Participants doing the task in the scanner made 3-4 runs of the task in each scanning session. In this setting, participants viewed a screen (30x48cm, 960x600px) through a mirror installed in the head coil. Stimulus sizes, fixation squares and radius were rescaled to compensate for the increased

distance subject-screen (laboratory facility: 60cm, scanner facility: 108cm approx.). In the laboratory facility, 21 participants completed a total of 4813 trials. 50 of them were excluded for wrong responses (no movement or wrong quadrant, 1.04%). Besides removing the wrong trials, outliers in each subject were removed using the interquartile range (IQR) method: Q1 – 1.5·IQR and Q3 + 1.5·IQR. 288 extra trials were removed (6.05%), leaving a total of 4475 trials. In the scanner facility, 6 participants completed a total of 2061 trials. 62 of them were excluded for wrong responses (3.01% -some subjects reported feeling sleepy due to the lying position inside the scanner) and 112 were outliers (5.6%). In total, the combined dataset consisted of 6362 trails in 27 subjects (70.33% obtained in the laboratory facility and 29.66% obtained in the scanner).

Measure of interference

In this thesis, I analyze behavioral vsWM in ring disposition. Previous studies showed that in this type of tasks, responses are systematically biased depending on the position of the visual space on which they appear (Girshick et al., 2011; Huttenlocher et al., 1991, 2004; Jastrow, 1892; Lipinski et al., 2010; Merchant et al., 2004; Pratte et al., 2017; Shin et al.,

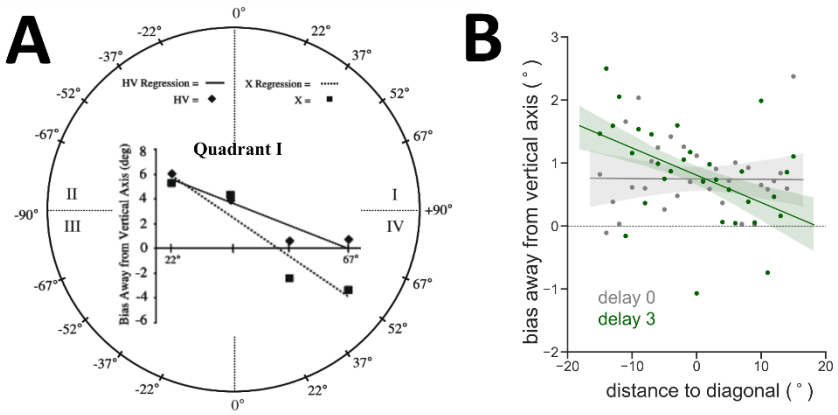

Figure 8 Axis effects on reports

A) Figure modified from Lipinsky et al. (2010). They observed reports biased away from the vertical axis (positive errors) when the targets were near this axis (±22° from vertical), and toward the vertical axis (negative errors) when the targets were near the horizontal axis (±67° from vertical). All quadrants showed a similar patter but, for clarity, just one is presented. **B)** Behavioral data of single-item trials of the paradigm *Topography of the WM circuit* showed a similar, with a general repulsion from the vertical axis (negative error) in the no-delay condition (grey) and an attractive effect towards the diagonal in the memory conditions (green line) that builds on top of the no-memory error. A mixed linear model revealed a significant interaction of axis effects with delay (linear mixed model, n= 3396 trials N= 18 subjects, dependent variable: attraction to vertical, *intercept* (β=-0.743, z=-2.790, ci=[-1.265, -0.221], p=0.005), *diagonal distance* (β=0.0, z=0.014, ci=[-0.027, 0.028] ,p=0.988), *delay* (β=-0.020, z=-0.346, ci=[-0.132, 0.092], p=0.730), interaction *delay*distance diagonal* (β=0.015, z=2.230, ci=[0.002, 0.028], p=0.026)

2017; Spencer & Hund, 2002; X. X. Wei & Stocker, 2015). Figure 8 shows an example of this effect in Lipinsky et al., (2010) and in one of my datasets. Although there is no consensus regarding the origin of these biases, plausible explanations have been proposed from the Bayesian framework (Girshick et al., 2011; X. X. Wei & Stocker, 2015), which state that previous information regarding the spatial disposition of presented stimuli generate a prior of more probable dispositions that could bias future perceptions and response. Taking into account that the stimuli of the vsWM tasks used are not completely uniformly distributed in the space (I avoided the cardinal axes to avoid any strategy based on using external references (Zelinski, 2016), like the screen frame), it was essential to use a measure of error that corrects for any possible systematic bias.

Figure 9 illustrates the importance of removing axis effects to correctly interpret interference effects. It shows that measuring interference as a signed error (target-response signed positive if the error goes towards the other item or signed negative if the error goes in the other direction) lead to misinterpretation of the data.

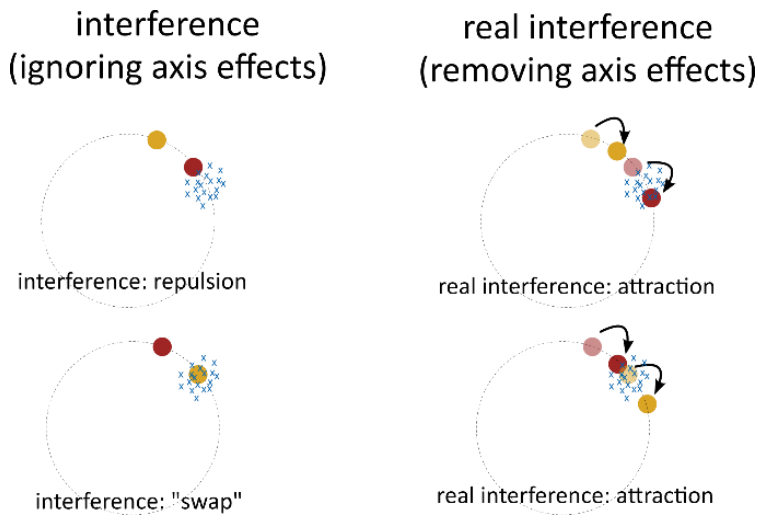

Figure 9 Error-misinterpretation due to cardinal axes effects
Measuring interference as a signed target–response measure leads to interpretation errors (left column) if axis effects are not removed (right). The red dot represents the target location; the yellow one represents the stimulus that interferes with it; and the blue crosses represent the responses. The first row shows an example of misinterpreting attractive effects as repulsion and the second row a misinterpretation of attractive effects as swap errors.

Previous works removed these biases by fitting non-linear functions and analyzing the residuals (Barbosa et al., 2020; Stein et al., 2020) or by some parameter of the fit as a measure of precision (Pratte et al., 2017). In this thesis, I developed a measure of interference based on the angular distance to the mean distributions of the errors when the reference item (the one to interfere with) is located cw or ccw to the target. Figure 10 illustrates this measure. Figure 10A shows the distribution of errors in the dataset of *Topography of the working memory circuit* when the non-target (NT) item is located cw or ccw. A difference between those distributions indicate interference, because the lack of interference would lead to completely overlapping distributions. Figure 10B shows the measure, which consists in aligning all the trials and then computing the angle error of each cw trial to the mean of the ccw distribution and dividing it by 2, and vice versa (Equation 1 and Equation 2).

$$interference_i^{cw} = \frac{err_i^{cw} - \overline{err^{ccw}}}{2} \qquad \textit{Equation 1}$$

$$interference_i^{ccw} = \frac{err_i^{ccw} - \overline{err^{cw}}}{2} \qquad \textit{Equation 2}$$

Figure 10C-D showed examples of surrogate data where the same interference attractive error of 5° was incorporated into a dataset with no effect of the axis (A) and to a dataset with a 10° repulsion from the vertical axis. In each plot, the interference error and the interference corrected by this method is shown. While standard measures of interference fail to calculate the real effect in data with systematic biases, the method that relies on the distance between the cw and ccw distributions successfully captured the real error in both scenarios.

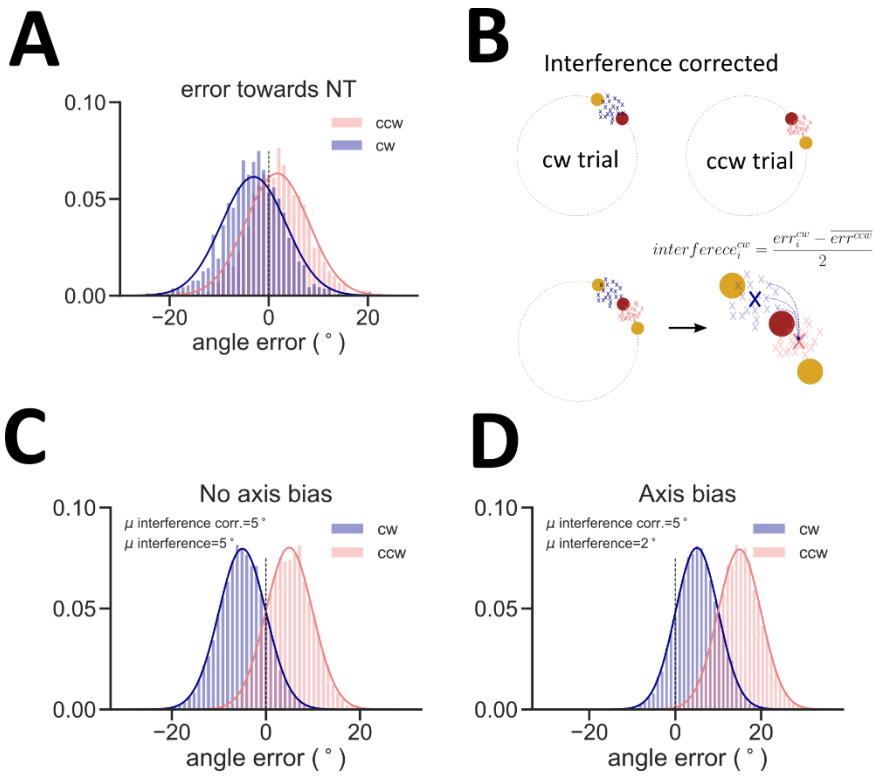

Figure 10 Measure of interference correcte.

A) Distribution of errors in the dataset with two simultaneous items in the dataset *Topography of the working memory circuit.* Significant difference between cw and ccw distribution is observed (t=18.13, p<0.001), indicating the presence of interference. **B)** Measure of corrected interference. Error in each cw trial is calculated by subtracting the mean of the ccw trials distribution and vice versa. **C)** Surrogated data with an interference error of 5°. Both the standard method of interference and the corrected method correctly estimated the real interference error when no systematic bias is imposed. **D)** Surrogated data with an interference error of 5°. Just the corrected method correctly estimated the real interference error when a systematic bias is imposed.

Computational modeling

Models of WM have been developed on different levels of abstraction (Durstewitz et al., 2000): purely psychological models, as the classical Baddeley-Hitch model (Baddeley, 2010; Baddeley & Hitch, 1974); highly abstract connectionist models, which neglect the temporal and spatial dynamics of neurons and synapses (Minami & Inui, 2001); firing rate models incorporating some biophysically meaningful time constants (Masse et al., 2019); and biophysically detailed models of spiking neurons (X.-J. Wang et al., 2004). In this work, I developed and extended a firing rate model formulation of the bump attractor model of Compte et al. (2000) to explain precision, interference, and distractor effects in vsWM.

The bump attractor model consists of a network of excitatory and inhibitory neurons (Wilson & Cowan, 1973) with spatial selectivity. Neurons selective for the to-be-remembered location present elevated activity during the delay period, forming a bump that diffuses during the delay period and it is maintained through reverberatory activity (Figure 11). Neurons with similar selectivity are strongly co-excited, while far away neurons are not. Since neurons with similar selectivity are modeled as close by neurons, a ring structure emerges (Figure 12A). This structure allows sufficient strong focal excitation to be kept in the form of a bump, even when the stimulus disappears.

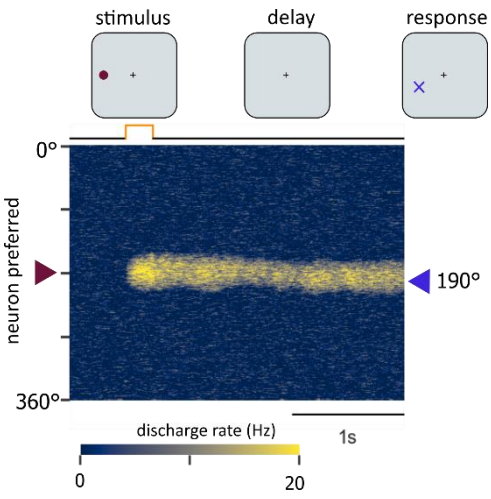

Figure 11 The bump attractor model
Example of a simulation of a vsWM trial with the bump attractor model. Activity of the neural network, in which the color represents the level of activity (blue: low activity; yellow: high activity). The bump of activity represents the evolution of the remembered location during the delay period. When the stimulus is presented (180°), neurons selective to this spatial location start to fire. When the stimulus is no longer present, activity is maintained thanks to reverberatory activity during the delay period. Due to noise fluctuations, the bump diffuses such that the final location of the bump is 10° away from the initial position of the stimulus.

In Chapter 1: *Topography of the working memory circuit,* I propose a plausible explanation for the topographical expansion of the model, both in the angular and the radial dimensions. In Chapter 2: *Distractor filtering in the working memory circuit*, I propose different control mechanisms to deal with distracting information in this model. One of them incorporates short-term synaptic plasticity to the model (Barbosa et al., 2020; Kilpatrick, 2018) and the other models the effects in distractor filtering of external cholinergic activation.

Network model in the angular dimension

I simulated a bump attractor network model in a firing-rate neuron formulation (Edin et al., 2009; Wimmer et al., 2014). The model consists of a population of excitatory neurons (N_E=512) connected to a population of inhibitory neurons (N_I=512). Neurons of both populations present selectivity for specific angles (θ_i for i=1..N), and neurons with similar selectivity are more strongly connected than those coding for distant locations. Since neurons with similar selectivity are located adjacent in the network, a ring structure emerges (Figure 12A). This connectivity follows a von Mises distribution (Equation 3, I_0 is a modified Bessel function of order 0), both for the excitatory and inhibitory populations. W_{EX} refers to the connectivity where the presynaptic unit is excitatory while in W_{IX}, the presynaptic unit is inhibitory. The combination of the connectivity profile

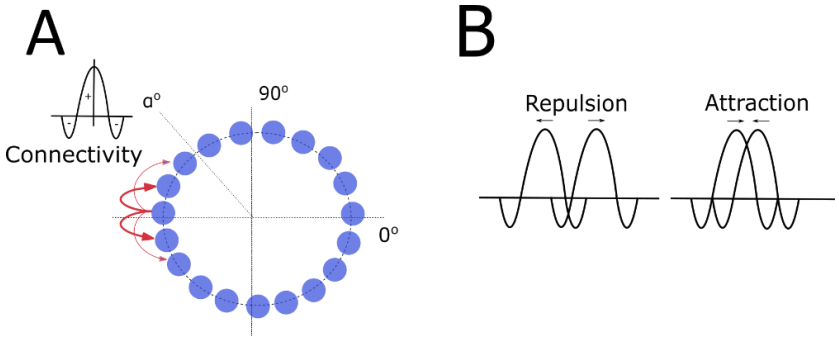

Figure 12 Connectivity of the bump attractor model

A) Ring structure of the model. Neurons with similar selectivity are strongly connected. The combination of the connectivity profile of the excitatory and the inhibitory population generate an overall Mexican hat connectivity, with inhibitory tails. **B)** The presence of inhibitory tails causes attractive or repulsive interference depending on the similarity distance. For close distances, attraction is expected. For far distances, repulsion is expected. In further distances, no interference is expected.

of excitatory and inhibitory connections generates an overall Mexican hat connectivity between excitatory neurons, with effectively inhibitory tails (Figure 12A). The attractor network allowed multiple memories to interfere during the delay period. The ring structure generates attractive effects for close-by memories and repulsive effects for distant ones (Figure 12B), as described in Almeida et al. (2015).

$$W_{ij}^{xy} = \frac{e^{k_x cos(\theta_i - \theta_j)}}{2\pi I_0(k_x)} \qquad x, y = E, I \qquad \textit{Equation 3}$$

I modeled two different eccentricities: radius 1 and radius 3. To model a loss of tuning with increasing eccentricity as the data revealed, I set a larger κ (1/κ is an analogous of variance in the normal distribution) for radius 1 (κ_E=300 and κ_I=30) compared to radius 3 (κ_E=225 and κ_I=15). Dynamical equations of the model describe how the rates of both populations decay in the absence of external current with a time constant τ (t_E=9, t_I=4). A white Gaussian noise (ξ) input to both populations, makes the bump diffuse randomly during the delay period (σ_E=0.5, σ_I=1.6) (Equation 4). The model transforms currents (I_n, n=excitatory or inhibitory) into rates (r) through a neural transfer function f(I)=0 for I≤0, f(I)=I^2 for 0<I<1, and f(I)=√(4I-3) for I≥1.

$$\tau_n \frac{dr_n}{dt} = -r_n + f(I_n) + \sigma_n \xi(t) \qquad n = E, I \qquad \textit{Equation 4}$$

The network couplings between excitatory and inhibitory neurons in the model (Equation 5 and Equation 6) are modulated by conductance parameters (G_{EE}=0.025, G_{IE}=0.01, G_{EI}=0.025, G_{II}=0.1). During stimulus presentation, an external current (I^0_n) is applied for 350ms to excitatory and inhibitory neurons with intensity peaking at the angular location of the stimulus according to a von Mises distribution (Equation 3), with κ_{stim}=150. If multiple stimuli are presented, external currents for each stimulus are summed.

$$I_E = I_E^0 + G_{EE} W_{EX} \cdot r_E - G_{IE} W_{IX} \cdot r_I \qquad \textit{Equation 5}$$

$$I_I = I_I^0 + G_{EI} W_{EX} \cdot r_E - G_{II} \bar{r}_I \qquad \textit{Equation 6}$$

The final readout -behavioral response (θ)- of the simulation was computed by extracting the population vector (Equation 7) of the activity of the excitatory population at the end of the delay, where r_j is the firing rate of the neuron j, and θ_j is its preferred selectivity, i is the imaginary unit (√-1) and arg the argument function in complex analysis. In the multi-item trials, more than one bump survived at the end of the delay. In this condition, I fitted a mixture of two von Mises functions and extracted the peak of each of the fits as independent readouts. Simulations where the bump died before the end of the delay or more bumps than stimuli appeared were discarded (7%). The measure of error was computed by subtracting the readout of the simulation from the initial position of the target. In the multi-item conditions, I measure interference towards the NT, which considers if the error with respect to the target was in the direction of the NT (positive: attraction) or not (negative: repulsion). As this network does not have cardinal axes effects, the measure of interference did not require correction.

$$\theta = arg(\sum_j r_j e^{i\theta_j}) \qquad \textit{Equation 7}$$

Network model in the radial dimension

This firing-rate formulation of the bump attractor model consisted of interconnected excitatory and inhibitory neurons (N_E=512, N_I=512) which presented selectivity for specific eccentricities (λ_i = iR/N for i=1..N with R as maximal radius) for a fixed angular coordinate. As a result, this was not a ring model. Neurons with similar selectivity were more strongly connected than those coding for distant locations. The connectivity strengths W_{ij} (Equation 8, where i and j are different neurons and x and y are excitatory and inhibitory neurons) followed a Gaussian distribution that changed with eccentricity as a function of $S_x(\lambda)$ (Equation 9).

$$W_{ij}^{xy} = \frac{e^{\frac{-(\lambda_i-\lambda_j)^2}{2S_x(\lambda_i)^2}}}{\sqrt{2\pi}S_x(\lambda_i)} \qquad x,y = E,I \qquad \textit{Equation 8}$$

This change followed an exponential function with different parameters for the excitatory and inhibitory profiles (a_E= 0.0033, b_E=1.7, c_E=0.05 and a_I=0.016, b_I=1.6, c_I=0.2).

$$S_x(\lambda) = a_x \cdot \lambda^{b_x} + c_x \qquad x, y = E, I \qquad \textit{Equation 9}$$

Stimulus presentation is modeled as an external current applied for 250ms to the excitatory population, with intensity peaking at the location of the stimulus according to a Gaussian distribution (Equation 8). Once this external input is no longer present, reverberatory activity maintains the information in the form of self-maintained selective elevated activity (a bump). The general equations that define how the rates evolve with time in both the excitatory and inhibitory populations are the same as for the angular model: Equation 4,Equation 5, Equation 6, with changes in the values of the parameters.

Time constants: τ_E=9, τ_I=4
White Gaussian noise: σ_E=0.8, σ_I=1.7
Conductances: G_{EE}=0.022, G_{IE}=0.01, G_{EI}=0.019, G_{II}=0.1

The memory readout of each simulation (L) was decoded from the firing rates of excitatory neurons at the end of the delay period using a population decoder for the radial dimension (Equation 10). To measure the attraction to fixation, I subtracted the final readout of the simulation to the initial position of the stimulus presentation.

$$L = \frac{\sum_i r_i \lambda_i}{\sum_i r_i} \qquad \textit{Equation 10}$$

Network model for distractor filtering

I implemented the bump attractor network model with short-term synaptic plasticity (STP) (Mongillo et al., 2008; Tsodyks et al., 1998), in line with the models of Kilpatrick (2018) or Barbosa et al. (2020). Figure 13 illustrates the dynamics of the STP mechanisms implemented. STP is modeled with two variables: x and u (Mongillo et al., 2008). The x variable denotes the fraction of resources that remain available after neurotransmitter depletion while the parameter u variable represents the fraction of available resources ready for use (release probability).

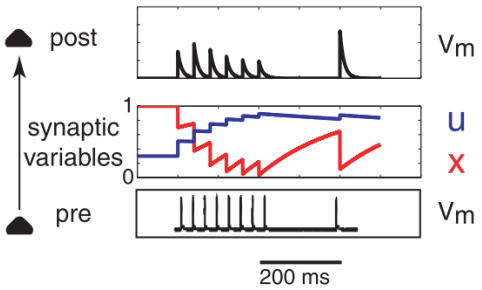

Figure 13 Example of postsynaptic response.
Figure form Mongillo et al. (2008). Example of the postsynaptic response to a train of presynaptic action potentials in the case of a facilitating connection. During the train, u -fraction of resources that remain available after neurotransmitter depletion (facilitation)- increases and x -fraction of available resources ready for use (depression)- decreases.

Upon a spike, some resources are used to produce the postsynaptic current, thus reducing x. This process mimics neurotransmitter depletion. The spike also increases u, mimicking calcium influx into the presynaptic terminal and its effects on release probability. When there is a spike, the amount of accumulated calcium increases by U(1-u) while the amount of available resources x is reduced by xu. In the rate version of the model, these dynamics are expressed as in Equation 11 and Equation 12. Between spikes, x and u recover to their baseline levels (x =1 and u = U) with time constants τ_x and τ_U, respectively. When $\tau_x > \tau_U$, the synapses mostly show short-term synaptic depression (STD) and when $\tau_U > \tau_x$, the dominant effect is short-term synaptic facilitation (STF).

$$\frac{du}{dt} = \frac{U-u}{\tau_u} + U(1-u)r_E \qquad \textit{Equation 11}$$

$$\frac{dx}{dt} = \frac{1-x}{\tau_x} - xur_E \qquad \textit{Equation 12}$$

STP mechanisms with the parameters τ_U=7000 τ_X=80 and U=0.4 are incorporated into the equations that describe the current of excitatory neurons (Equation 13), so synaptic efficacy is modulated by the product of x and u. STP mechanisms are not incorporated into the equations that describe the current of excitatory neurons (Equation 6). The time constant parameters of the STP induce an initial STD followed by STF.

The current of each population is modulated by conductance parameters (G_{EE}=0.016, G_{IE}=0.012, G_{EI}=0.015, G_{II}=0.007), time constants (τ_E=60, τ_I=10), uncorrelated Gaussian white noise (σ_E=0.06, σ_I=0.04) and κ values (κ_E=100, κ_I=1.5, κ_{stim}=20) for the connectivity (Equation 3).

$$I_E = I_E^0 + G_{EE}W_E \cdot (xur_E) - G_{EI}W_I \cdot r_I \qquad \textit{Equation 13}$$

With these parameters, the network maintains reverberatory bump attractors following a transient tuned input to the network. The elevated firing during stimulus presentation induced a large drop in x (STD) that takes a certain time to recover. Then, as $\tau_u > \tau_x$, STF prevails, and the synaptic efficacy is increased gradually until it reaches a stable firing. The combination of synaptic depression and facilitation at the onset of persistent activity in this network model induces a dip in the firing rate of memory selective neurons in the early delay period (Figure 14).

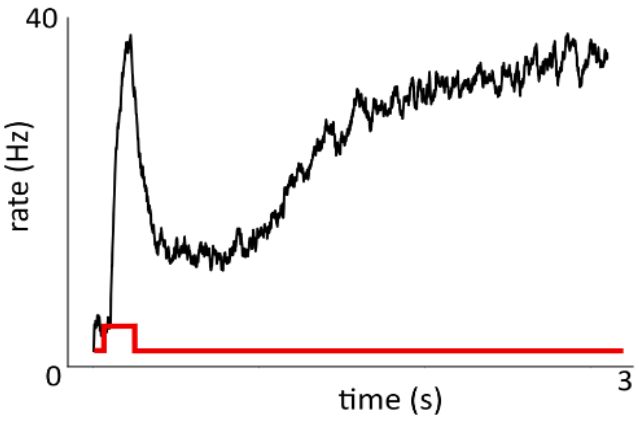

Figure 14 Example of rate dynamic. The combination of rate dynamics and the temporal parameters used for STP induces a dip in the firing rate in the early delay period that slowly recovers until it reaches a stable state.

In this model, I also incorporated a control mechanism to differentiate relevant from irrelevant information (target-distractor). In these attractor networks, small increases in nonspecific external currents can make the network transition from nonstable regimes, in which uniform baseline activity is robust to transient stimulation, and bistable regimes, in which transient inputs destabilize baseline activity and trigger bump solutions. This gives a dynamical mechanism to control whether the network stores or not inputs through changes in external currents (I^0_E). I externally modulated I^0_E to switch between stable resting and spatially structured states (I^0_E= 0.5 and sI^0_E = 1.2, respectively) depending on whether the presented stimulus was the target or the distractor (Equation 14). The network always started in resting state, and it switched to the spatially structured state whenever a target was presented.

$$I_E^0 = I_E(t) + I_E(\theta_i, t) \quad \textit{Equation 14}$$

Network model of NB stimulation (rate model)

I developed a bump attractor network model in a firing-rate formulation as in the *Model of the angular dimension*. I did not model different eccentricities, and the connectivity followed Equation 3 with κ_E=45 and κ_i=0.3 for excitatory and inhibitory connections respectively. Visual stimuli were modeled as an external current applied for 100ms to excitatory neurons during stimulus presentation, with intensity peaking at the location of the stimulus according to a Von misses distribution with κ_{stim}=40. In addition, I simulated an "expectation signal" concomitant with the stimulus presentation as an additional non-specific input to the excitatory network with strength (as in Equation 14). This represented an internal signal that predicted the regularly timed stimulation presentation events in the task, and it was relevant to simulate the emergence of phantom bumps when the stimulus was not presented (Figure 70).

The general equations that define how the rates evolve with time in both the excitatory and inhibitory populations are the same as for the angular model (Equation 4, Equation 5 and Equation 6) with changes in the values of the parameters.

Time constants: τ_E=20, τ_I=10
White Gaussian noise: σ_E=9.2, σ_I=6.6

I modeled the remember-first or remember-second conditions of the experiment with slightly different connectivity parameters and input (Compte et al., 2000) to reflect top-down influences in the specific blocks of these conditions. Specifically, I modulated G_{EE} (1st=0.068, 2nd= 0.064), G_{II} (1st=0.13, 2nd=0.01196), G_{EI} (1st=0.13, 2nd=0.1482), and G_{IE} (1st=0.042, 2nd= 0.045), and I^0_E (1st=-3.5, 2nd=-2).

To model the Nucleus Basalis stimulation (ON condition), I assumed that release of acetylcholine (Ach) in PFC would result in blockade of hyperpolarizing intrinsic currents in excitatory neurons, so I increased the excitability of this population through an increase of I^0_E (ΔI^0_E, 1st=3.55, 2nd=2.05). For all conditions I^0_I was set to 0.5.

Network model of NB stimulation (spiking model)

I simulated a bump attractor network model using the spiking network model of Hansel & Mato (2013), implemented in *Brian1*. The network consists of 20,000 integrate and fire neurons (16,000 excitatory) sparsely connected but maintaining the network architecture of the ring model by means of a translationally invariant Gaussian probability of connection that depends on their distance on the ring (Hansel & Mato, 2013). Excitatory synapses onto excitatory neurons displayed short-term synaptic plasticity, with effective facilitation dynamics as described above. A detailed explanation of both the single neuron dynamics and the connectivity of the network can be found in Hansel & Mato (2013). I slightly modified the parameters of the synaptic interactions and the external current to reproduce the regime of the experiment with a single bump.

Modified parameters of the synaptic interactions from the original in Hansel & Mato (2013):

G_{EE} (AMPA) = 490 mV/ms

G_{EE} (NMDA) = 490.64 mV/ms

U=0.04

Modified parameters of the external current from the original in Hansel & Mato (2013):

$stim_E$ (stimulus input) = 0.24 mV.

I_I^b (baseline external current inhibitory) = 1.54 mV

The total time of the simulations was 7 seconds, with a single stimulus onset at second 2. The stimulus presentation duration was 1 second. For the computation of the bump diffusion as well as for quantifying the tuning curves, the individual stimulus could be presented in 25 different positions covering the space 0°-360° (separation of 14.4°). I modeled NB stimulation as a general increase in the excitability of prefrontal neurons by slightly increasing the external input to excitatory units in the network. In the OFF condition, I_E^b (baseline external current excitatory) was set to 0.0 while in ON trials it was set to 0.5mV.

Magnetic resonance imaging (MRI)

Data acquisition

I scanned 6 participants for multiple sessions (8h approx.). Two of the participants were scanned more times to pilot the different tasks (11 and 17 sessions respectively). The remaining 4 participants were scanned for at least 4 sessions, each of them lasting 1.5 – 2h. In those sessions, 5 different types of tasks were used: *polar retinotopy*, *eccentricity retinotopy*, *encoding task*, *working memory localizer,* and *working memory task.* The first four tasks were presented in runs of 5min and the *working memory task* was presented in runs of 10min. After each task, participants rested 1-5min.

The *eccentricity retinotopy* and the *polar retinotopy* were used to define the visual ROI (region of interest). The first one (Figure 15A) consisted in a 5min task with cycles of 30s where a flickering ring (flickering rate=4Hz, max. radius=20cm) expanded from the fixation point to the periphery (*eccentricity retinotopy out*) or contracted towards fixation (*eccentricity retinotopy in*). The *polar retinotopy* task (Figure 15B), consisted in a 5min task where a wedge of 45 degrees (flickering rate=4Hz, radius=20cm) continuously rotated cw or ccw (*polar retinotopy cw* and *polar retinotopy ccw*). A complete rotation (cycle) took 30s. For both tasks, participants had to maintain fixation and indicate with a click when the fixation point changed color, which happened four times per cycle.

The *encoding task* was initially used for training the IEM (Sprague et al., 2014) but it was later discarded as the IEM was trained in the *working memory task*. The *encoding task* (Figure 15C) consisted in a 5 min task where participants had to remember the spatial location of a black dot (radius=1cm, presentation period=500ms) located at a radius of 16cm during a delay period of 3s. During the delay period, a flickering circular checkerboard (flickering rate=4Hz, radius=5.4cm was displayed on top of it (not completely centered). At the end of the delay period, subjects had to report whether another black dot was at the same position or if instead it was slightly displaced.

The w*orking memory localizer task* (Figure 15D) consisted in a 5 min task where participants were initially presented for 13s with a flickering band (6Hz, internal radius=14cm, external radius=18cm) of 90 degrees covering

a whole quadrant. After the 13s and with the band still flickering, two consecutive WM trials were displayed. In these trials, participants had to remember the spatial position of a black (radius=1cm, presentation period=500ms) dot appearing in the quadrant for 3 seconds (appeared at 16cm from fixation). After the delay period, participants had to report whether a green dot (same dimensions) was located upper or lower

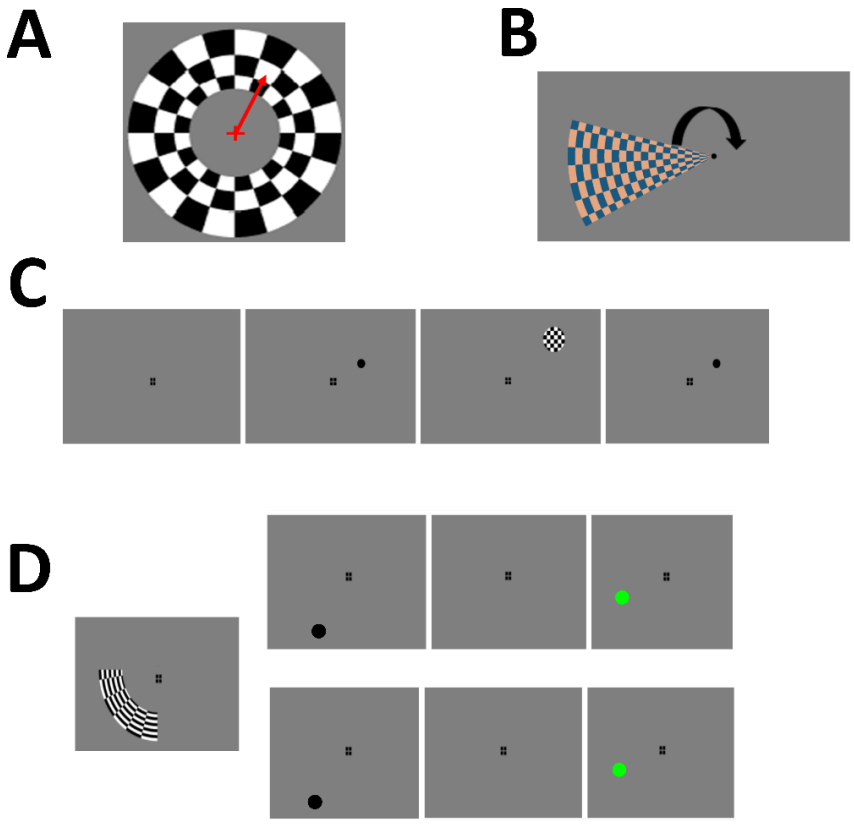

Figure 15 Spatial mapping and localizer tasks.

A) *Eccentricity retinotopy task*. Cycles of 30s of a flickering ring expanding from fixation (out, showed here) or contracting from periphery (in) This task was used to identify visual regions. **B)** *Polar retinotopy task*. Rotating wedge flickering at 4Hz (cw rotation in the figure). A whole rotation of 360° was accomplished in 30s. This task was also used to identify visual regions **C)** *Encoding task*. This task was initially used to train the inverted encoding model (IEM) as in Sprague et al., (2014). This task was later discarded as the IEM was trained in the actual WM task. Participants had to remember the location of a black dot, and, during the delay period, a flickering circle (4Hz) was presented on top. At the end of the delay period (3s), participants had to judge whether a black dot was at the same position or not. **D)** *Working memory localizer task*. A flickering band (6Hz) of 90 degrees covered a whole quadrant alone for 13s. Then, on top of it, two consecutive WM trials were displayed were participants had to remember the spatial position of a black dot appearing in the quadrant for 3 seconds and report whether a following green dot was located upper or lower respect the black dot (two lower trials displayed). This task was used to define the parietal and frontal regions.

respect the black dot (two lower trials displayed. Flickering is constant during the whole trial). This task was used to define the parietal and frontal regions. The *working memory task* is described in the Results section of WM and distractor filtering (Figure 43).

All six participants were scanned on a 3 T General Electrics MRI system equipped with a using a 32-channel receiver only head coil at the *MR Centrum* of the Karolinska Institutet, Stockholm. A head holder was used to prevent head movements, and earplugs were used to attenuate the scanner noise. As participants had to provide reports, they were taught to use a scanner-compatible mouse. Anatomical high-resolution three-dimensional T1-weighted scans were acquired for each participant (TR=6.4ms, TE=2.8ms, 256 slices, flip angle=90°, voxel size =1 x 1 x 1mm).

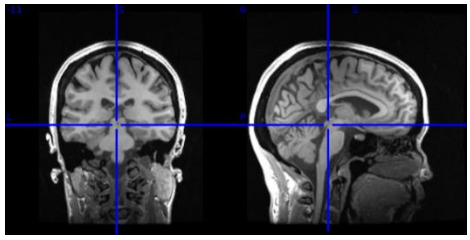

Figure 16 Anatomical scan.
Example of the T1 high- definition anatomical scan of one of the participant (explicit consent was obtained to publish this picture).

Functional magnetic resonance imaging (fMRI) data was acquired using a gradient echo planar imaging (EPI) pulse imaging pulse sequence. For the retinotopy tasks (*eccentricity retinotopy* and the polar *retinotopy*) I used the following scanning parameters: TR=1600ms, TE=30ms, 30 slices, flip angle=130°, voxel size =2 x 2 x 3mm. For the rest (*encoding task*, *working memory localizer* and *working memory task*), I used the following scanning parameters: TR=2335ms, TE=30ms, 46 slices, flip angle=90°, voxel size =2 x 2 x 3mm). The associated task paradigms were programmed in with the python library *Psychopy* (Peirce et al., 2019), and were initiated by a trigger sent by the scanner. Participants viewed a screen through a mirror installed in the head coil. The screen (30x48cm, 960x600px) was located at 105cm from the mirror. As the *working memory task* was previously ran in the laboratory facility, stimulus size and radius were doubled (stimulus size of r=0.5cm to r=1cm and radius of presentation of 8cm to 16cm) to compensate the increased distance subject-screen and maintain the visual angle constant (laboratory facility: 60cm approx., scanner facility: 110cm approx.). The same correction was applied for the *encoding task* and the *working memory localizer*.

Data preprocessing

Preprocessing was carried out either with SPM8 (Statistical Parametric Mapping, http://www.fil.ion.ucl.ac.uk/spm) or with FreeSurfer (FreeSurfer Software Suite, https://surfer.nmr.mgh.harvard.edu). SPM8 was used to preprocess volume fMRI data that did not need to be transform into surface (*encoding task* and the *working memory task*) while FreeSurfer was used to preprocess volume fMRI data that was transformed into surface to extract the ROIs (*eccentricity retinotopy, polar retinotopy* and *working memory localizer*).

Volume preprocessing

1- Realignment

Although participants are instructed not to move, there are always unavoidable head movements, and thus the data becomes corrupted with motion related artifacts. Thus, head motion correction should be performed on all fMRI data (Figure 17). Motion correction was performed by registering all volumes to a reference volume (middle volume of the whole dataset) via a rigid-body transformation (Park et al., 2019).

Figure 17 Realignment
Schematic view of head motion correction. Figure from Park et al. (2019).

2- Slice timing

Slice timing correction is performed to correct the time differences at which each slice is acquired. To do so, slice timing correction uses interpolation, which causes a temporal smoothing effect, with a possible loss of information (Park et al., 2019). Slice timing is not recommended if the TR is short (<1 s) (Bijsterbosch et al., 2017) but I decided to apply it as the TR was 2.335s.

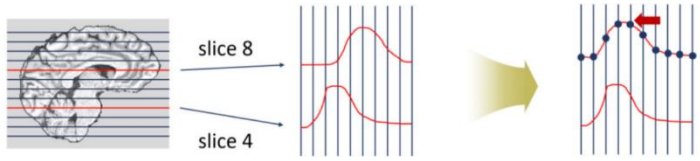

Figure 18 Slice timing correction.
Example of slice timing correction where the time of the signal evoked at slice 8 is shifted toward that of slice 4 to match the starting time. Figure from Park et al. (2019).

3- Corregistration to T1

Corregistration is the process of aligning images from the same subject, for example an anatomical and a functional image (similar process as the realignment, finding x,y, and z parameters for translation and rotation). Low-resolution fMRI data is corregistered onto high-resolution preprocessedT1-weighted structural MRI data of the same subject via a rigid-body transformation (Park et al., 2019).

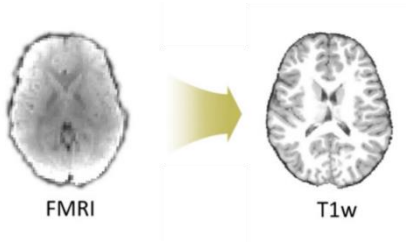

Figure 19 Corregistration
fMRI data is registered onto structural MRI data of the same subject via a rigid-body transformation. Figure modified from Park et al. (2019).

4- Corregistration to a single-subject fMRI template

The analysis was conducted in the subject space, so no normalization to standard spaces like the Montreal Neurological Institute (MNI) was applied. Instead, all fMRI data from the same subject was resliced to a common fMRI template image (one for each subject, coming from the *encoding task*). This step was needed because the ROIs were obtained from fMRI data in the surface, and they were later transformed into volume using the same fMRI template. I did not directly corregister to the fMRI template because adding the intermediate step of corregistering to a high-resolution MRI signal (T1 weighted) increased the accuracy of the final corregistration compared to a direct one.

5- Spatial Smoothing

Spatial smoothing (Figure 20) is achieved by calculating the weighted average over neighboring voxels using a Gaussian. The full width at half maximum (FWHM) of the kernel was 4 x 4 x 4mm (Worsley & Friston, 1995). Spatial smoothing offers the advantage of reducing noise, but it also can lower the intensity of the signal (Park et al., 2019).

Figure 20 Spatial Smoothing
Figure from Park et al. (2019).

Surface preprocessing and ROI definition

I applied surface preprocessing using FreeSurfer for the *eccentricity retinotopy*, *polar retinotopy* and *working memory localizer*. The preprocessing included *Realignment, Corregistration* (to the surface space of the anatomical image) and *Spatial Smoothing* (5 x 5 x 5mm). In this case, I decided not to apply *Slice timing*, because the TR for the *eccentricity retinotopy* and polar *retinotopy* was shorter (TR=1600ms).

I used FreeSurfer to create a retinotopic mapping of the visual areas. Retinotopic areas of the visual system were identified by fitting a linear model of the expected pattern of activation to the measured BOLD signal in the *eccentricity retinotopy* and the *polar retinotopy* tasks (Arcaro et al., 2009; Brewer et al., 2005; Dougherty et al., 2003; Engel et al., 1994).

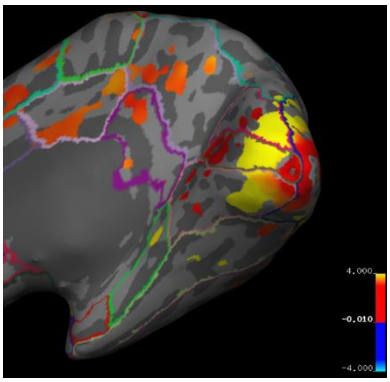

Figure 21 Retinotopic mapping
Result of retinotopic mapping in the right hemisphere surface of one subject.

In the case of the *polar retinotopy* task, for example, where a rotating wedge is presented, for a voxel representing the 0° location, the measured BOLD signal can be model by a cosine (starting at maximal response). The other polar locations can be modeled with cosines of same frequency but different phases. As successive visual areas alternate with a mirror or non-mirror representation of the visual field, borders between these areas can be delineated when a reversal of the map is detected. To obtain a better mapping, the model combines the expected pattern of BOLD signal produced by an expanding ring (Figure 15A) with the expected pattern produced by a rotating wedge (Figure 15B). An example of the retinotopic analysis is presented in Figure 21. For the final ROI definition of V1, the retinotopic analysis was combined with the FreeSurfer Atlas.

The ROI definition of parietal and frontal areas was accomplished through the *working memory localizer*. Although retinotopic mapping also was useful for the ROI definition in parietal regions, I wanted ROIs to be defined based on the BOLD signal of a WM task, to select voxels that present elevated activity during the delay period. To do so, after surface

preprocessing the fMRI data from the *working memory localizer,* I created a design matrix with two events: *baseline* and *WM*. The event *WM* had a duration of 3s, coinciding with the delay period in each trail (Figure 15D). The analysis was also run in FreeSurfer and the significant voxels were combined with a FreeSurfer Atlas (Figure 22) of cortical labeling (Desikan–Killiany–Tourville protocol) to extract the ROIs in inferior parietal, *superior parietal* and *superior frontal* cortex (Klein & Tourville, 2012). The significance level was varied from subject to subject to get similar ROIs in each subject. Once the ROIs were identified in the surface space, they were defined using the FreeSurfer visualization tool *TkSurfer*. Figure 23 shows an example of the final ROIs in the surface space in visual, parietal, and frontal cortex for one subject. Finally, the masks were transformed to volume space using the same fMRI template as for the volume preprocessing corregistration (*volume preprocessing: 4- Corregistration to fMRI template*).

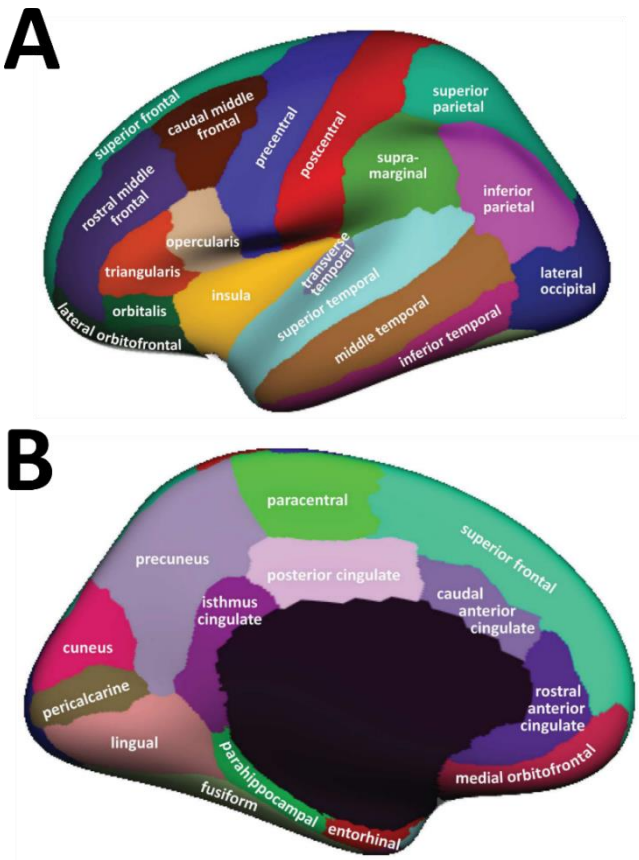

Figure 22 FreeSurface atlas of cortical regions
Lateral **(A)** and medial **(B)** views of the inflated cortical surface. I combined *inferior parietal*, *superior parietal,* and *superior frontal* with the WM localizer task to extract parietal and frontal ROIs. Figure from Lein & Tourville (2012).

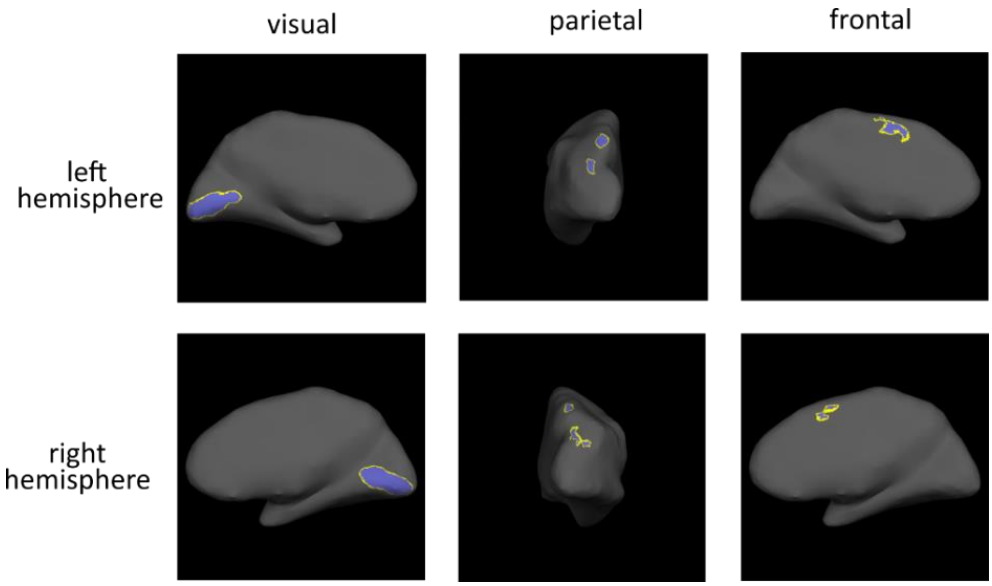

Figure 23 ROI in the surface space
Visual, parietal, and frontal ROIs for one subject in the right and left hemispheres

The specific codes an extra documentation for the volume and surface preprocessing of the data (either in SPM or FreeSurfer), the retinotopy analysis, the *baseline vs WM* analysis, and the instructions to create the masks can be found in the following GitHub repositories:

https://github.com/davidbestue/Preprocessing-functional-images

https://github.com/davidbestue/Retinotopy

https://github.com/davidbestue/WM-localizer

Inverted encoding model (IEM)

I implemented an inverted spatial encoding model to reconstruct WM content using the pattern of BOLD signal during each TR of the task (Brouwer & Heeger, 2009; Ester et al., 2013; Sprague et al., 2014). This method assumes the BOLD signal reflects an approximately linear combination of neural responses (Figure 24A) that changes in each voxel (Figure 24B). It also assumes voxel stability in the response, so the BOLD pattern for the training dataset is maintained in the testing dataset. Previous works used an independent task to train the model (Ester et al., 2015; Sprague et al., 2014, 2016) while others used also TRs of the actual WM task (Hallenbeck et al., 2021; Lorenc et al., 2018; Rademaker et al., 2019). Although I started by training the model on an independent task, the results presented in this thesis are obtained with the IEM trained in the same WM task.

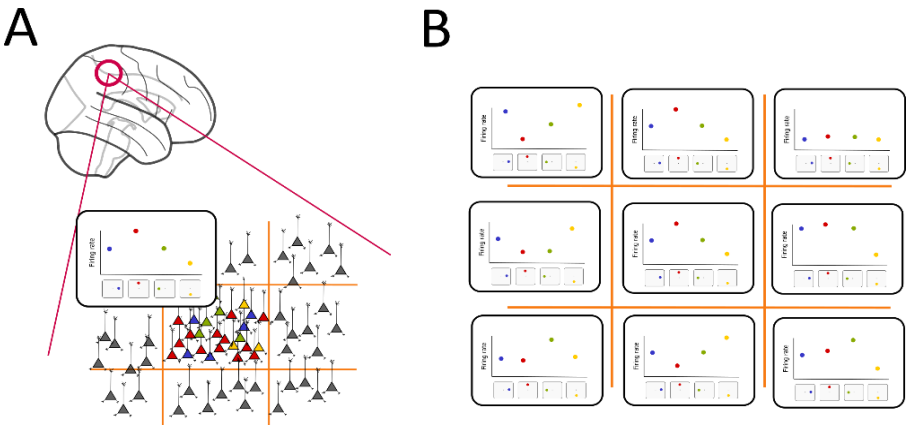

Figure 24 Voxel's relation to spatial selectivity
A) Each voxel contains neurons with different spatial selectivity, so the final firing response is a combination of them. **B)** Selectivity changes from voxel to voxel due to neural heterogeneity.

As the stimuli in the WM task were disposed in a ring, we changed the typical structure of the model (Sprague et al., (2014): a grid of 36 information channels) to a ring structure with 36 information channels (Figure 25A) gaining precision in the angular dimension (peaks at 5°-355° every 10°). The model response followed Equation 15, where r is the distance from the stimulus position to the center of the channel function,

and s is a size constant which corresponds to the distance from the channel function center at which the function reaches zero (s=48.518°).

$$f(r) = (0.5 + 0.5cos\frac{2\pi r}{s})^7 \qquad \textit{Equation 15}$$

Channels are spaced by 10°. Figure 25B shows the response of two adjacent channels to stimulus located at 175° and 185°. Figure 25C shows the channel response to the 4 axis positions (0°, 90°, 180° and 270°).

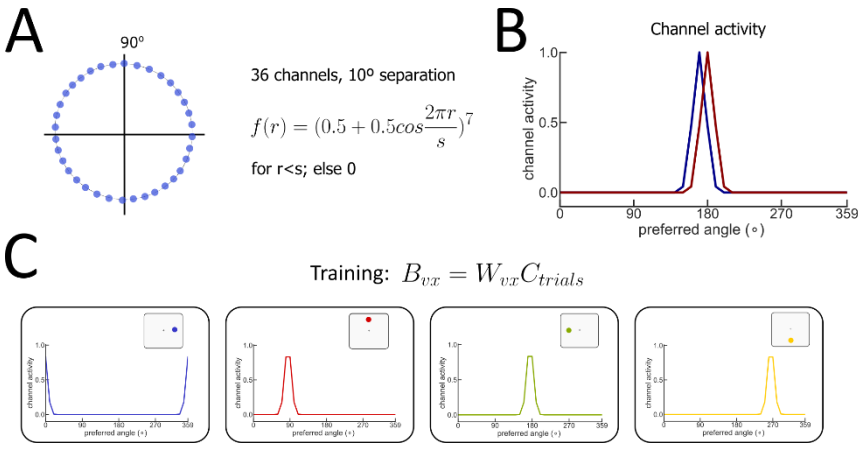

Figure 25 Training the IEM

A) The IEM disposed 36 information channels along the angular dimension, separated by 10°. **B)** Response of the model (36 channels) to stimuli separated by 10° (175° and 185°). **C)** Channel responses to stimuli located at the axis.

For training the IEM, I measured the BOLD signal of each voxel and modeled it as a linear combination of the 36 channels (Equation 16), where B_{vx} is the observed BOLD signal in each voxel on each trial (*i* voxels x *n* trials), C_{trials} is the predicted response of each channel to the stimulus location in each trial (*j* channels x *n* trials) and W_{vx} are the weights of each channel to explain BOLD signal (*i* voxels x *j* channels).

$$B_{vx} = W_{vx}C_{trials} \qquad \textit{Equation 16}$$

The contribution of each channel (W_{vx}) to the observed response in each voxel is estimated through an ordinary least-squares regression (Equation 17). This step is performed on each voxel individually.

$$\widehat{W_{vx}} = B_{vx}C_{trials}^{T}(C_{trials}C_{trials}^{T})^{-1} \quad \textit{Equation 17}$$

Once the weight matrix is estimated, we can apply it to reconstruct the WM content from the BOLD signal of the WM task in the ROIs (B_{ROI}) (Equation 18, Figure 26A). The resulting estimated channel responses (C_{rec}) (Equation 19) reflect the response of each channel that is most likely to have given rise to the observed pattern of activation across all voxels within an ROI, given the observed BOLD signal.

$$B_{ROI} = \widehat{W_{vx}}C_{rec} \quad \textit{Equation 18}$$

$$\widehat{C_{rec}} = (\widehat{W_{vx}^{T}}\widehat{W_{vx}})^{-1}\widehat{W_{vx}^{T}}B_{ROI} \quad \textit{Equation 19}$$

As the reconstruction in each trial was very noisy as multiple stimuli were presented (3 targets and 3 distractors), the estimated channel reconstruction in each trial was rotated so the channel coding for the target position was always at 180° (Figure 26B, up). To avoid reconstructing the distractor information instead of the target information, training and testing was done aligning to the target or the distractor that was presented alone in the quadrant (*Methods-Paradigm distractor filtering*). As both training and testing data came from the WM task, I cross-validated each reconstruction. To do so, I trained the model in a single scanning run and tested it in the remaining ones (leave-one-run-out). I did that for all the scanning runs and I computed the final average reconstruction. Figure 26C shows the mean reconstruction for one subject in the condition *order 1 -long TDOA*. The x axis shows the different TRs and the y-axis the angle. Reconstruction around 180° is stronger (brighter colors mean stronger channel signal) because target information was aligned to it. The grey triangle illustrates the target presentation time, the red one shows the distractor presentation time and the yellow one the response.

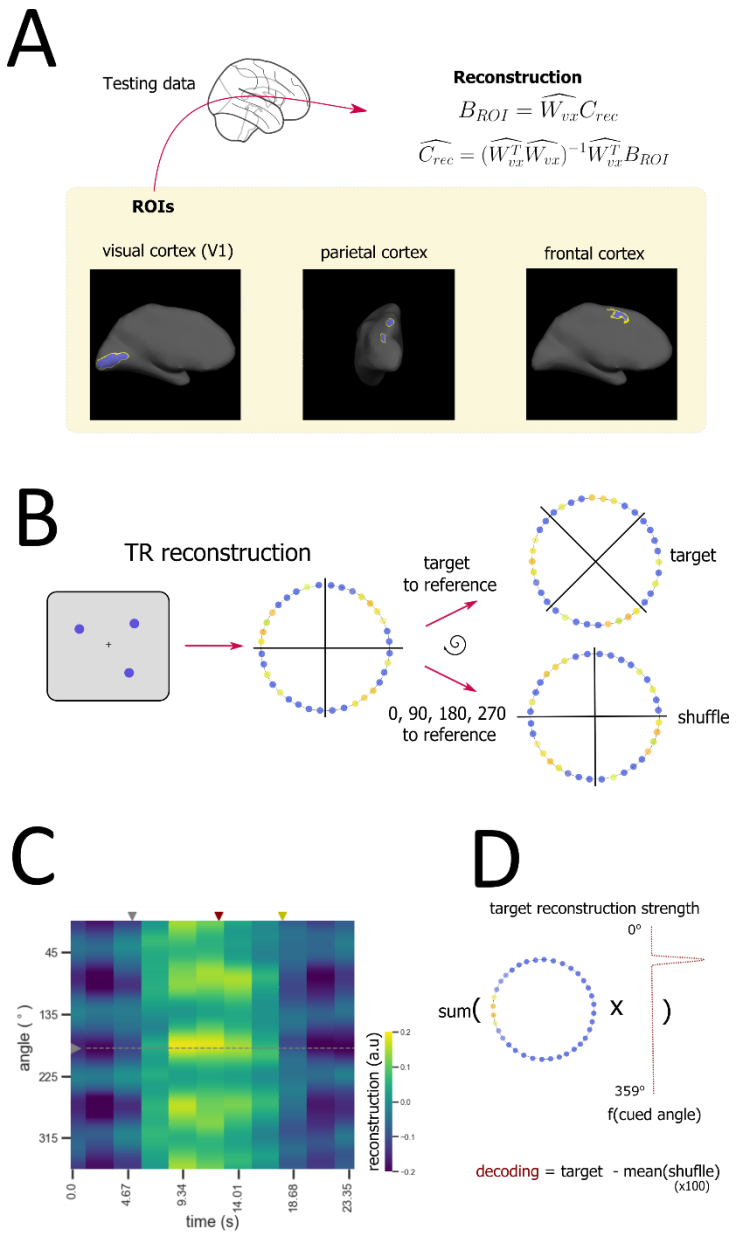

Figure 26 Testing the IEM

A) Reconstructing WM content in different ROIs using the trained IEM. **B)** Schematic view of the reconstructed channel activity for a certain TR in a trial. Reconstructions were aligned to a reference (180°) to calculate the mean signal or randomly aligned to an axis position for a posterior subtraction of baseline activity. **C)** Mean reconstruction of aligned targets for the *order 1 -long TDOA* condition. **D)** Decoding strength for a certain TR is the weighted sum with a high-resolution model (720 channels) and the decoding value is calculated by subtracting the activity in the shuffled reconstructions for the position 180°.

To avoid confounding baseline activity with WM content, I also computed the reconstruction but, instead of rotating the channel reconstruction so the channel coding for the target position was at 180°, I rotated the channel reconstruction so one of the 4 axis positions (0°, 90°, 180°, 270°) was at 180º. I used this method instead of a pure random assignment of angles because, as previously exposed, up to 6 visual stimuli that could elicit a WM related response were presented and a complete shuffle would capture WM related response as baseline activity, compromising the signal to noise ratio. For each condition, I ran 100 of these reconstructions (shuffles) by randomly assigning which axis position was aligned to 180° (Figure 26B, down).

Finally, for each TR, I calculated the reconstruction strength for both the target and the shuffles as the weighted sum of each channel activity with a high-resolution model. To do so, I created a high-resolution model with 720 channels instead of 36 (same Equation 15). For each channel, I multiplied the reconstructed activity by the expected signal of the high-resolution model when a stimulus was presented at the location mapped by channel. Then, I summed all these 36 arrays of 720 values to get a single array of 720 values (Figure 26D), which is the high-resolution reconstruction of a TR (Figure 26C). The final decoding value is obtained by getting the activity at 180° (channel 360 of 720) and subtracting the

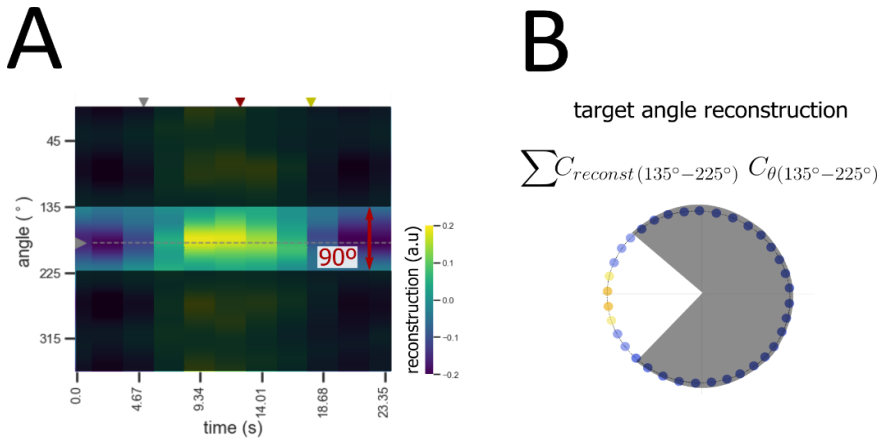

Figure 27 Reconstruction of target angle

A) As three different targets were simultaneously presented, I restricted the decoding to 90° centered to the target to avoid contamination. **B)** A standard population vector was computed to extract the reconstructed angle.

mean activity at the same location for the shuffles. Subtracting the mean decoding of the shuffles is needed to avoid any possible decoding coming from averaging noise.

Alternatively, I also used another method to decode the exact angle (Figure 27). To do so, after creating the high-resolution reconstruction of each TR, I restricted the analysis to the range -45° to 45° around the reference (Figure 27A) and computed the population vector in it (Figure 27B). The rest of the visual space was neglected to avoid confounding the decoding with the one of the targets or the distractors presented in other quadrants.

Electrophysiology

Electrophysiological datasets of monkeys performing vsWM tasks were used in each chapter of the results section. For the chapter *Topography of the working memory circuit*, I analyzed the Dataset1 and for the chapter *Distractor filtering in the working memory circuit*, I analyzed the Dataset2 and the Dataset 3.

Dataset 1: O'Scalaidhe & Goldman-Rakic, unpublished

Unpublished neurophysiological dataset from O'Scalaidhe and Goldman-Rakic, referenced in Arnsten (2013), where two monkeys performed an ODR task. Unfortunately, this dataset lacked the behavioral responses of the monkey, but I could still analyze relevant aspects of the neuronal tuning. The task consisted in remembering the spatial location of a square during a delay period of 3s and then reporting its location with a saccade. The stimuli could be located at 24 different spatial locations (Figure 28) divided in three eccentricities -8 locations in each- (Arnsten, 2013). 516 neurons were individually recorded from the PFC when doing the task. I used a linear regression with the firing rate during the presentation period as the dependent variable and each angular location as an independent regressor to get the neurons with a significant angle tuning. From the total of 516 neurons, 153 presented angular selectivity. I then assigned a preferred eccentricity based on the maximum firing rate in each of the 24 positions, so the 153 angle-tuned neurons were split in the three eccentricities (53 in radius 1, 64 in radius 2 and 43 in

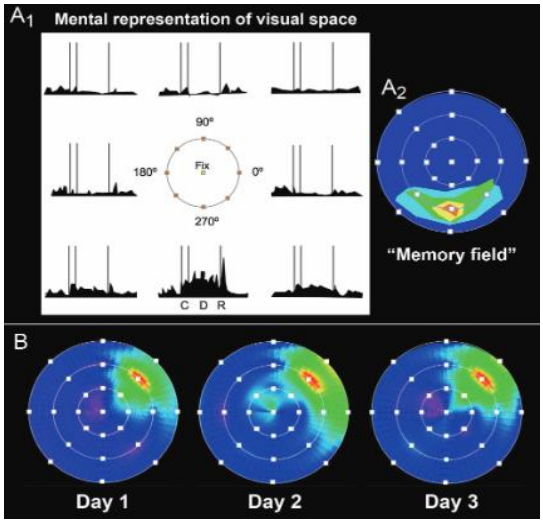

Figure 28 O'Scalaidhe & Goldman-Rakic dataset
Figure 4 in Arnsten (2013). This figure shows the disposition of the presented stimuli in three different radii. This figure was ised in a Goldman-Rakic presentation for Yale undergraduates and it was used with her consent.

radius 3). Once each neuron was assigned to an eccentricity, I calculated the z-scored tuning curve for neurons in each eccentricity. To quantify the differences between tuning curves, I calculated the circular standard deviation of the tuning curve of each neuron.

Dataset 2: Suzuki & Gottlieb (2013)

Neurophysiological dataset from Suzuki & Gottlieb, (2013) where two monkeys performed a modified version of the ODR task with distractors (Figure 29A). After a variable fixation period (300-800ms), monkeys were presented with a 100-ms flash of a peripheral target and, after a 1600ms delay period, made an eye movement to the remembered location. On two-thirds of the trials, a 100ms distractor was flashed during the delay at a randomly selected TDOA of 100ms, 200ms, 300ms or 900ms. Distractor's location was also randomly selected to be near (45° angular separation) or far from the target location (135° or 180° separation). The target and distractor were identical in appearance and duration, so that monkeys had to remember the location of the first stimulus and suppress the subsequent distractor. Both monkeys had near perfect performance for the long TDOA condition (900ms) at near or far locations (Figure 29B).

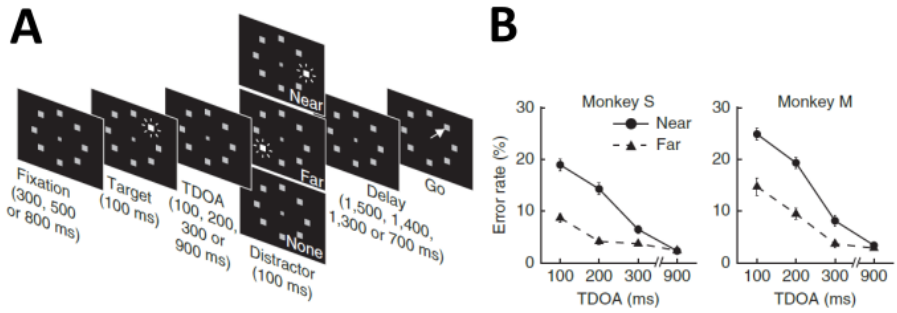

Figure 29 Behavioral results Suzuki & Gottlieb (2013).
Figure from Suzuki & Gottlieb (2013). **A)** Task structure. An array of eight placeholders remained continuously on the screen and a trial began with a variable period of central fixation. This was followed by a 100ms flash indicating the target location. After a variable TDOA, a distractor flashed after target presentation. The distractor was identical to the target in appearance and duration but appeared at either a near-target or far locations. After an additional delay (bringing the total delay period to 1600ms) the fixation point disappeared (Go) and monkeys were rewarded for making a saccade to the target location. **B)** Performance for each monkey as a function of distractor distance and TDOA (mean and sem across all recording sessions, n=89 sessions in monkey S, n=47 sessions in monkey M)

However, errors became increasingly more common as distractors became more similar to the target in both time and space. Statistical analysis revealed significant effects of distance and TDOA, and a significant interaction, such that the largest fraction of errors was found for near-target, short TDOA conditions (Suzuki & Gottlieb, 2013).

Neural recordings were collected from 77 spatially tuned neurons in the dorsolateral prefrontal cortex (dlPFC) (51 in monkey S, 26 in monkey M) and from 59 neurons in lateral intraparietal area (LIP) (38 in monkey S, 21 in monkey M, not analyzed in this thesis). All the neurons selected had spatial receptive fields as determined by preliminary testing with the memory-guided saccade task (further details in Suzuki & Gottlieb, (2013), online Methods).

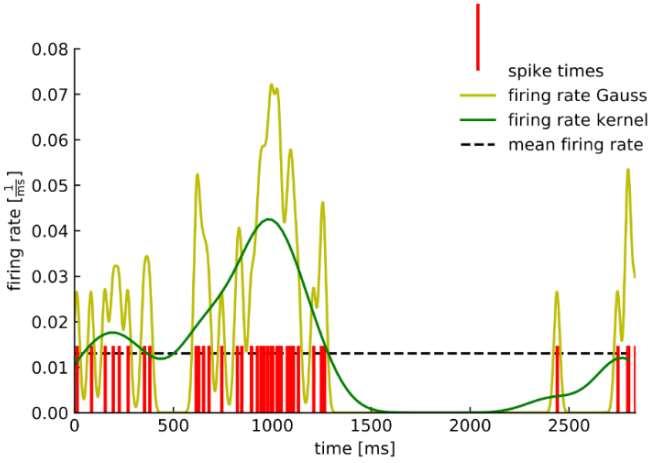

Figure 30 Single neuron analysis

Analysis of an individual neuron spike train (red) in a trial (2600ms). Mean firing rate is shown with a dotted black line in 1/ms units. The green shows a convoluted firing rate using the default kernel of the library *Elephant* with a sampling period of 25ms. The yellow line shows the convoluted firing rate with the specificities of Suzuki & Gottlieb, (2013): Gaussian kernel with sampling rate of 2ms and sliding window of 15ms.

From the spike times of each neuron (Figure 30, red), I computed the convoluted firing rate. As in Suzuki & Gottlieb, (2013), I used a sampling period of 2ms and a Gaussian kernel with a sliding window of 15ms (Figure 30, yellow). The analysis was computed with the python library *Elephant* (Denker et al., 2018).

To measure distractor responses, the firing rates of each neuron were also normalized by dividing by the peak target–evoked response (T in receptive field, no-distractor). The replication of the analysis gave equivalent results as the original paper (Figure 31).

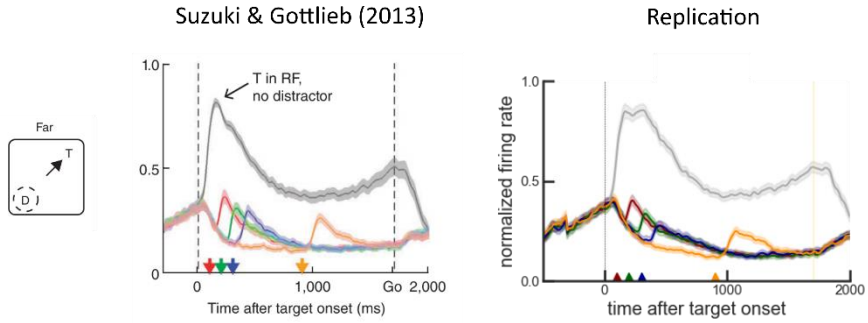

Figure 31 Replication analysis of Suzuki & Gottlieb, (2013)
Replication of the neural responses in dlPFC for the condition where the distractor was in the receptive field (RF) and the target was located far. Colored lines showed the average normalized firing rates in PFC aligned on the onset of the target. The grey line is the peak target-evoked response, and the different colors represent the neural response to the distractor for the different TDOA conditions. On the left, I illustrate the original plot from Suzuki & Gottlieb (2013) and, on the right, my replication analysis of the same data, both showing equivalent results. Differences may be originated by the libraries used for the neural analysis.

I evaluated the stability of the memory circuit by training and testing a decoder of the remembered angle in the rates of the recorded neurons every 100ms of the task (Murray, Bernacchia, et al., 2017; Stokes et al., 2013). As a decoder, I used a multivariate regression with the *sin* and the *cos* of the target location (y_{i1} = sin(θ) and y_{i2} =cos(θ)) as dependent variables and the firing rate as independent variable (Equation 20 and Equation 21).

$$y_{i1} = [\beta_{01} + \beta_{11} rate_{i1}] + \varepsilon_{i1} \qquad \textit{Equation 20}$$

$$y_{i2} = [\beta_{02} + \beta_{12} rate_{i1}] + \varepsilon_{i2} \qquad \textit{Equation 21}$$

Dataset 3: Qi et al. (2021)

I modeled the behavioral and electrophysiological results of Qi et al., (2021). In this paper, the authors implanted two adult male rhesus monkeys (*Macaca mulatta*) with two 20-mm-diameter recording cylinders (arrays of two to four microelectrodes in the cylinder). One cylinder was located over the areas 8a and 46 of the dlPFC -used for the study- and the other over the PPC -not used for the study-. The anatomic location of electrode penetrations was determined based on MR imaging (Figure 32). Once the cylinders were implanted, a second surgery was performed to implant the stimulating electrode in the center of the anterior portion of the Nucleus Basalis of Meynert (NB).

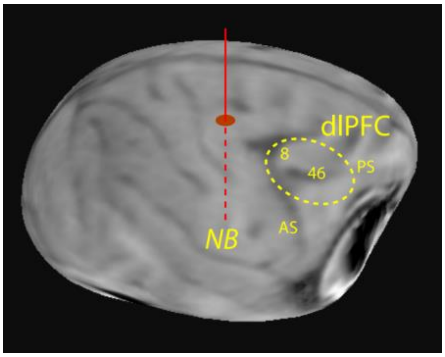

Figure 32 MR implant location
Anatomical MR scan representing the cortical region sampled with neurophysiological recordings (yellow dotted area) and the approximate location of the NB-stimulation electrode (red). Figure from Qi et al. (2021).

The monkeys were trained to perform a variation of the ODR task, where two visual stimuli appeared in sequence. The monkeys had to remember and make an eye movement to the location of either the first or the second visual stimulus depending on the color of the fixation point (Qi et al., 2021). The task (Figure 33A) was administered in blocks, so some blocks had prospective distractors (*Remember 1st*: remember the first stimulus and ignore the second) and other blocks had retrospective distractors (*Remember 2nd:* ignore the first and remember the second stimulus). The monkeys were trained to saccade to the location of the remembered visual stimulus according to the color of fixation point (white/blue). Once the monkey fixated at the center of the screen for 1s, two white squares were displayed sequentially for 0.5s, with a 1s stimulus onset asynchrony (SOA) (-1s or +1s TDOA depending on whether the distractor was the first or the second stimulus presented). Each stimulus was displayed at one of eight possible locations arranged along a circular ring (spaced 45 degrees).

The angular distance between the stimuli could then be 0°, 45°, 90°, or 180°. Two ‘‘null’’ conditions were included, in which either the first or the second stimulus was omitted. After the presentation of the second stimulus, a delay period of 1s was introduced -delay2-. The monkeys were rewarded with juice after making a correct saccade. Deviating gaze beyond fixation window led to the immediate termination of the trial without reward. Intermittent stimulation of the NB was applied for 15s at 80 pulses per second, followed by approximately 45s with no stimulation. The stimulation was applied during the inter-trial interval (IT) (Figure 33B). Extra details regarding the task structure can be found in the original article (Qi et al., 2021).

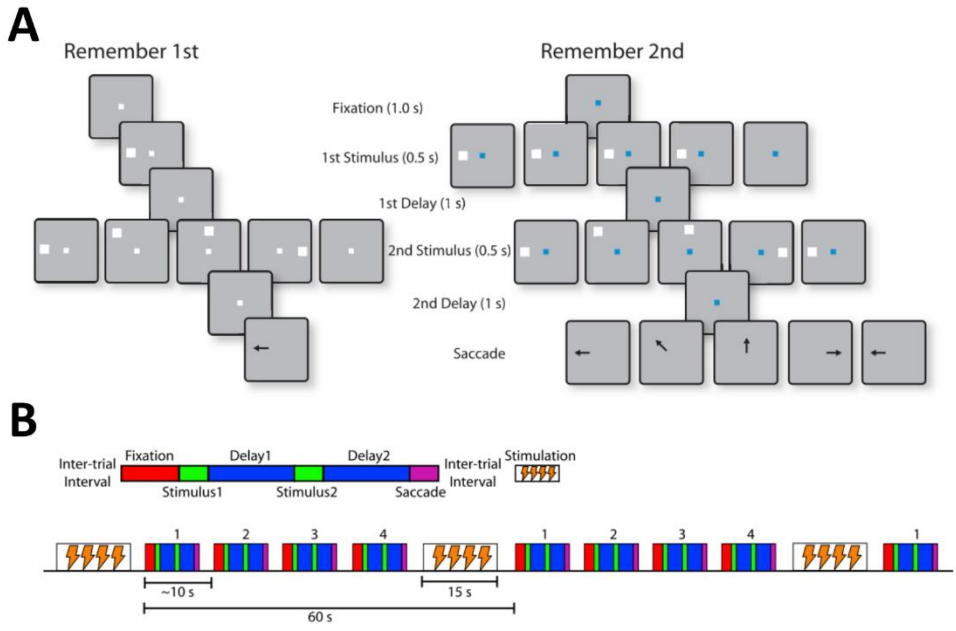

Figure 33 vsWM task of Qi et al. (2021)

Figure from Qi et al. (2021). **A)** Schematic view of the two types of trials of the behavioral task: *Remember 1st* and *Remember 2nd*. The type of trial was imposed by the color of the fixation point. White fixation point indicated the monkey had to remember the first visual stimulus presented, and a blue fixation point indicated it had to remember the second one. The duration of the different periods of the trials are indicated in the figure. At the end of the trial, the fixation point turned off, and the monkey had to perform a saccade toward the remembered location to receive a liquid reward. **B)** Top: schematic diagram of a single trial of the task. Bottom: Blocks of the same type of trial (*Remember 1st* or *Remember 2nd*) were presented successively, separated by an ITI. NB stimulation, when delivered, always occurs during the ITI. Successive trials, each lasting approximately 10 s, were followed by 15 s of stimulation.

4.Results

4.1. Topography of the working memory circuit

In this chapter, I addressed goals **(1)** and **(2)** through the investigation of the topographical features of the vsWM circuit both in the angular and the radial dimensions in behavioral experiments. First, I tested a fundamental point of the SRT **(1)**, which is the predicted interference between encoding and maintenance processes based on the shared neural circuit (Fischer & Whitney, 2014; W. J. Harrison & Bays, 2018; Teng & Kravitz, 2019): "as a neural circuit is shared, perceptual responses based on immediate reaction upon incoming information should present the same topographically-based biases as responses in delayed paradigms that require WM". I used the bump attractor model **(2)** to propose a mechanistic explanation for the observed topographically-based biases in the angular dimension (Figure 34, red) and I also provided electrophysiological evidence supporting it (*Methods-Electrophysiology*, Dataset 1). Besides explaining topographically-based biases in the angular dimension, I also developed a computational model **(2)** to explain topographically-based biases in the radial dimension (Figure 34, blue). The latter model explained mislocalizations in the radial dimension towards the fixation point, which is a well-documented behavioral effect (Müsseler et al., 1999; Osaka, 1977; Sheth & Shimojo, 2001; Townsend, 1973) lacking a detailed mechanistic explanation.

Altogether, in this chapter I developed computational models for both the angular and radial dimensions, and I provided topographically-based behavioral, modeling, and electrophysiological evidence for prefrontal attractor networks being responsible for the final WM readout.

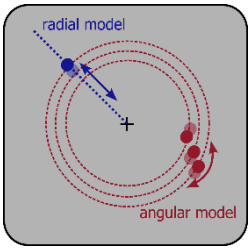

Figure 34. Topography and modeling
Schematic view of the dimensions covered by each of the developed computational model. One explains WM maintenance in the angular dimension (red) and the other in the radial dimension (blue). The solid blue dot represents a target stimulus and the transparent blue a drifted response in the radial dimension. Solid red dots represent target stimuli at different eccentricities and the transparent ones, the diffused responses (cw or ccw) in the angular dimension.

Angular dimension

To test the convergence of the memory circuit and the encoding circuit proposed by the SRT, I first investigated how the precision of sensory-guided and memory-guided reports depended on the eccentricity of the reported location. To do so, I ran a behavioral experiment with 18 participants where single or multiple stimuli needed to be remembered at different eccentricities with variable delay periods (Figure 35, *Methods-Paradigms and analysis*).

I first analyzed the precision of responses in single-item trials (Figure 36). I considered both the angular distance (measured as degrees of azimuthal angle) and the Euclidean distance (measured in cm) between the response endpoint and the target locations. As expected, absolute Euclidean errors grew with eccentricity but the evidence was not strong for them being larger for memory (delay 3s) than for no-memory (delay 0 s) trials (Figure 36A), mixed linear model, N=18, n=3275, β *eccentricity*=-0.091, z=14.044, ci=[0.078,0.103], p<0.001; β *delay*=0.052 , z=1.537 , ci=[-0.014, 0.118], p=0.124; β *delay X eccentricity*=0.002, z=0.594, ci=[-0.004, 0.008], p=0.553).

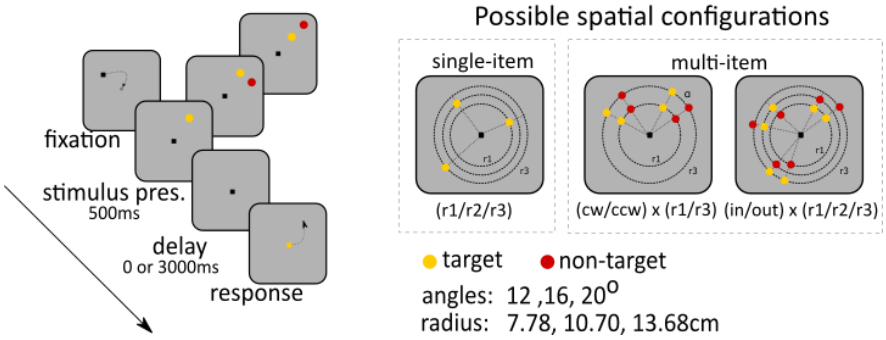

Figure 35 vsWM task to study the topography of the WM circuit.
Trial structure (left). Participants had to fixate at the center of the screen (both looking and dragging the cursor) to start the trial. Stimuli were presented for 500ms. When multiple stimuli were presented, each had a different color. A variable delay of 0s or 3s was introduced before the fixation point changed to one of the colors of the stimulus presented, indicating the subject had the report its location. There were three different possible spatial configurations of trials (right): Single-item trials (right, single item), multi-item angular (right, multi-item, left) and multi-item radial (right, mulit-item, right). The last ones were not not analyzed in this thesis. Stimuli were always presented in one of three radii (7.78cm, 10.7cm, 13.68cm) and, in multi-item angular trials, they were located at either radius1 (7.78cm) or radius 3 (13.68cm) and separated by one of three angular distances (12°, 16°, 20º). More details in *Methods-Paradigms and analysis*.

Because I wanted to interpret these results on the basis of the bump attractor model, I was specifically interested in the modulation by eccentricity of precision in the angular domain. I wondered if changes of absolute Euclidean precision with eccentricity could merely reflect a radial scale change on otherwise identical precision in the angular dimension. This would correspond to the absence of a main effect of eccentricity when applying the mixed linear model to the angular error. Instead, I found a significant interaction between eccentricity and delay (Figure 36B, N=18, n=3275, β *delay*=0.730, z=4.995, ci=[0.443, 1.016], p<0.001; β *eccentricity*=-0.028, z=-0.998, ci=[-0.083, 0.027], p=0.318; β *delay x eccentricity*=-0.038, z=-2.825, ci=[-0.064, -0.012], p=0.005). The interaction reflected a decreasing absolute error with eccentricity in the

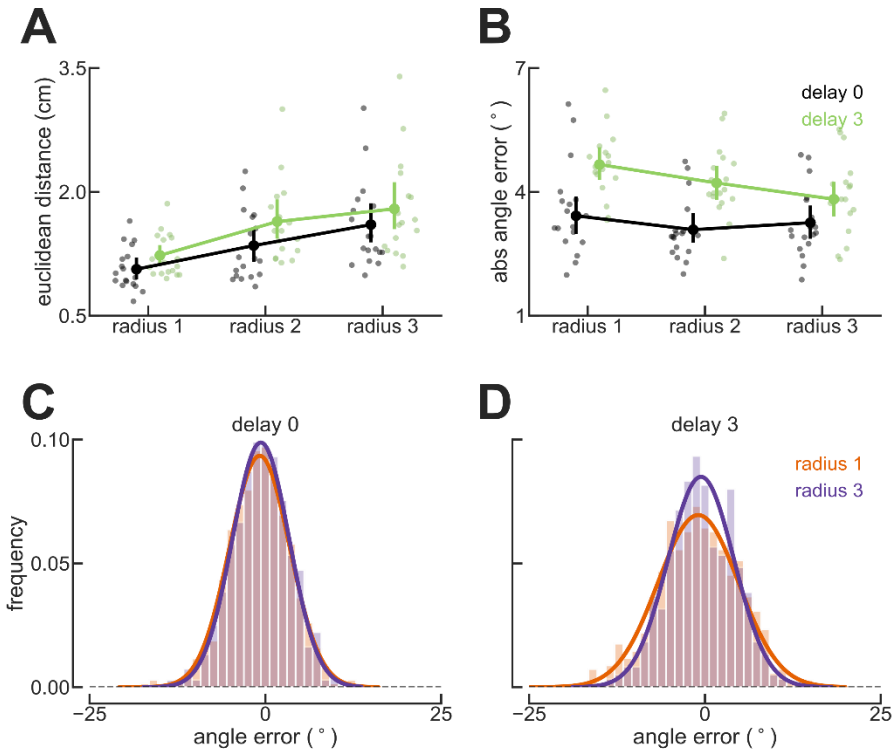

Figure 36 Single-item trials show memory-specific increase of angular precision with eccentricity.
A) Euclidean distance grows with eccentricity and shows a non-significant trend with memory (delay 0 vs. delay 3). Each small point is the mean error of each subject at the indicated radius (random x-scatter introduced only for visualization purposes). Line-connected dots show population means and 95% bootstrapped cis. **B)** Same for angular errors reveals interaction between eccentricity and delay for absolute angular error. **C)** Distribution of angular errors in the delay 0 condition shows similar accuracy for radius 1 and radius 3, as estimated through std of Gaussian fits (4.2° (sem=0.29°) and 3.8° (sem=0.23°), Levene's test W=0.86, p=0.35). **D)** Angular errors in the delay 3 condition reveal higher dispersion of responses in radius 1 than radius 3 (5.4° (sem=0.19°) and 4.6° (sem=0.25°), Levene's test W=17.98, p<0.001).

memory condition (n=1595 trials in the delay 3 condition, β *eccentricity*=-0.142, z=-4.384, ci= [-0.206, -0.079], p<0.001) and no effect of eccentricity in the delay 0 condition (n=1680 in the delay 0 condition, β *eccentricity*=-0.028, z=-1.143, ci=[-0.075, 0.020], p=0.253). In the no-memory condition (delay 0), I observed no difference in the distribution of errors at different radii (Figure 36C, Levene's test W=0.86, p=0.35), consistent with absolute Euclidean errors reflecting a radial scaling of identical angular error distributions in the absence of memory requirements. However, in the memory condition (delay 3s), I observed an unexpected increase in angular accuracy with radius (Figure 36D, Levene's test W=17.98, p<0.001). I thus found that the Euclidean precision of memory reports decayed with eccentricity (Figure 36A), but the angular precision increased with eccentricity (Figure 36B). This last effect was specific to memory processes because it was not verified in no-memory trials (significant interaction). This different scaling of angular accuracy with eccentricity for the no-memory and memory conditions points towards neural circuitries with different topographical arrangements for encoding and WM.

A model based on a retinotopic map with RF sizes increasing linearly with eccentricity, consistent with the neurophysiology of visual cortex, would explain the results in the no-memory condition. For the task, this would be conceptually modeled as identical concentric rings of neurons for each of the radii considered, with identical angular tuning of neurons in all rings. However, the decrease in response errors with eccentricity for the memory condition points towards a different topography for the circuit responsible for memory maintenance. To accommodate these findings, the memory model must incorporate eccentricity-dependent changes in the parameters of the concentric ring models. In order to gain more intuition on the topography of the memory-maintenance model, I analyzed the pattern of interference between multiple stimuli (Almeida et al., 2015; Nassar et al., 2018) located at different eccentricities in our multi-item trials (Figure 35, right). I regressed the angular report errors around the target stimulus (*Methods-Paradigms and analysis*) against the delay duration and the eccentricity and distance between presented stimuli, and I found a significant triple interaction (mixed linear model, n=2291, β *eccentricity x delay x distance*=-0.011, z=-2.556, ci=[-0.020, -0.003], p=0.011). This indicated the existence of differences in how multiple memory items interfere at different eccentricities for memory and no-memory trials.

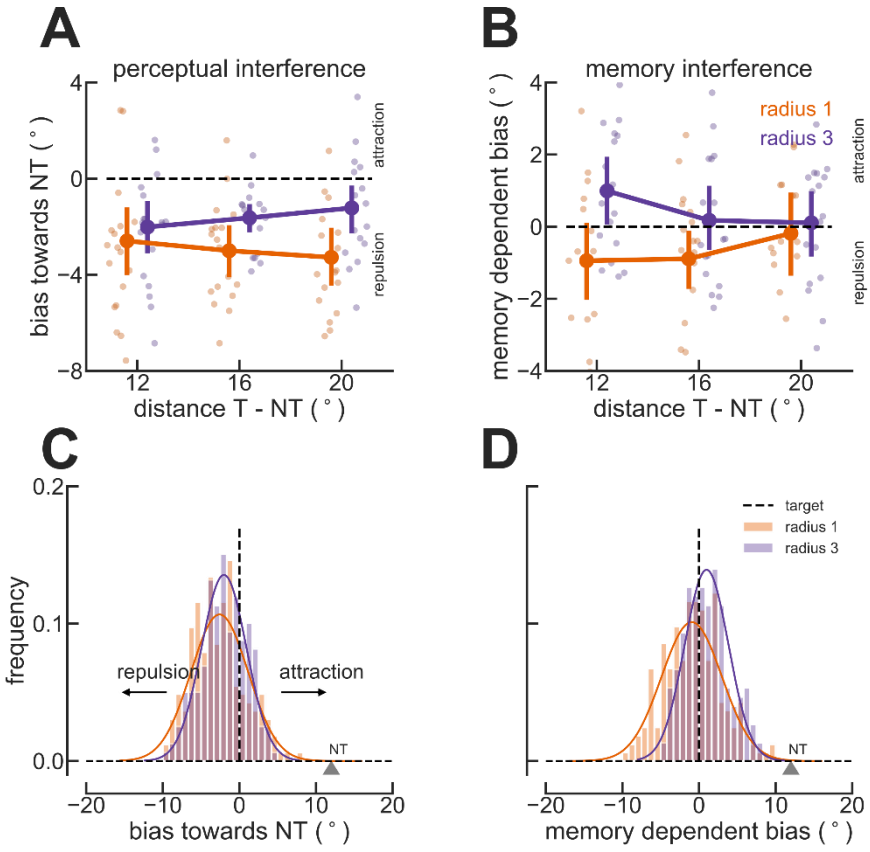

Figure 37 Repulsive interference in multi-item trials has a memory component that depends qualitatively on eccentricity.

A) Mean response bias towards the location of the NT item in no-memory trials (delay 0) for different absolute distances between stimuli (3 discrete values on x-axis, additional horizontal scatter was introduced for better visualization) and for two different radii (orange and blue). Positive (negative) interference effect indicates attraction (repulsion) between items. Line-connected dots show population means and 95% ci. Data show overall perceptual repulsion between items, larger for small than large eccentricities, and opposing effects with inter-item distance at different radii. **B)** Same for memory-dependent interference effect, defined as additional mean bias towards NT item in delay-3 trials compared to the mean of corresponding delay-0 trials. Memory-dependent interference was repulsive for small radius and attractive for the large one. Interference diminished for both radii with target-NT distance. **C)** Distributions of errors towards the location of the NT item for delay-0 and 12° inter-item separation trials (n=201 for radius 1, n=191 for radius 3, collapsed across 18 participants) show repulsive bias for both radii. **D)** Distributions of memory-dependent interference biases in delay-3 and 12° target-NT distance trials show an overall repulsive bias in radius 1 and an overall attractive bias for radius 3.

I analyzed separately the pattern of interferences in no-memory trials, reflecting interferences that occur with minimal involvement of the memory maintenance circuit. In this case, I observed that the interference between target and NT stimuli was consistently repulsive (Figure 37A). This overall repulsive effect showed an eccentricity-dependent pattern: the larger the distance between target and NT, the stronger the repulsive effect in radius 1 and the weaker in radius 3 (mixed linear model, N=18, n=1169, β *intercept*=0.388, z=0.267, ci=[-2.457, 3.233], p=0.789; β *eccentricity*=-0.275, z=-2.161, ci=[-0.524, -0.026], p=0.031; β *distance*=-0.033, z=-3.82, ci=[-0.497, -0.160], p<0.001; β *eccentricity x distance*=0.031, z=3.991, ci=[-0.016 -0.046], p<0.001). Repulsive effects at the perceptual level have been described before (Fritsche et al., 2017; Gibson & Radner, 1937; O'Toole & Wenderoth, 1977; Stein et al., 2020) and probably reflect biases affecting inputs to the memory areas (Bliss & D'Esposito, 2017; Stein et al., 2020).

Because my interest was in characterizing memory processes, I analyzed biases in memory trials that added to these perceptual repulsions. To this end, I subtracted the mean bias in no-memory trials from response errors in memory trials, in order to remove the perceptual effects from those that belong to memory processes. I observed that there was still an overall repulsive effect for radius 1 but an overall attractive effect for radius 3 (Figure 37B). Furthermore, there was an interaction between distance target - NT and eccentricity: The larger the distance between target and NT, the weaker the repulsive effect in radius 1 and the weaker the attractive effect in radius 3 (mixed linear model, N=18, n=1122, β *intercept*=0.388, z=0.267, ci=[-2.457, 3.233], p=0.789; β *eccentricity*=0.739, z=4.346, ci=[0.406, 1.073], p<0.001; β *distance*=0.361, z=3.153, ci=[0.136, 0.585], p=0.002; β *eccentricity x distance*=-0.035, z=-3.342, ci=[-0.055, -0.14], p=0.001). Figure 37C-D show the distribution of errors of a representative condition (distance target-NT of 12°) to illustrate the patterns of perceptual and memory interferences at different radii. In line with the accuracy results in single-item trials presented above, the distinct pattern of eccentricity-dependent interferences between simultaneous items in no-memory and memory trials points towards different neuronal circuitries being responsible for perception and memory. Specifically, the data suggest that the perceptual circuit generates systematic repulsive biases between simultaneously presented

items, which then serve as input to a separate memory circuit that builds an attractive bias on top with distinct topography.

I hypothesized that eccentricity-dependent changes in the connectivity parameters of the bump attractor model could account for the effects of single-item and multi-item trials for two main reasons. First, delay-dependent loss of memory precision occurs naturally in bump attractor models because activity bumps are susceptible to noise fluctuations, and this susceptibility is modulated by connectivity parameters (Compte et al., 2000). Second, attractive and repulsive regimes between simultaneously memorized items, as seen behaviorally, was previously modeled (Almeida et al., 2015; Nassar et al., 2018) via changes in the connectivity parameters in these types of models. I thus tested how the general widening of the connectivity in these models affected the precision of an individual memory, and the pattern of interference between two simultaneous memories, in order to compare with our experimental findings. I built a ring attractor network of excitatory and inhibitory rate-model neurons, recurrently connected through translationally-invariant Gaussian-like connectivity profiles (*Methods-Network model in the angular dimension*), which was able to maintain up to two stable memory bumps during the delay period through reverberatory activity. Based on my hypothesis, I modeled angular memory at different eccentricities (radius 1 and radius 3) by introducing changes in the connectivity parameters leading to wider connectivity profiles for larger eccentricities. As a result of this different connectivity, model simulations produced a broader bump (on angular terms) in radius 3 compared to radius 1 (Figure 38A).

I first analyzed memory precision in simulations with a single memory bump for each of the two networks (radius 1 and radius 3). I reasoned that a broader activity bump in radius 3 would be more effective in averaging out noise fluctuations and explain the unexpected increase of angular accuracy from radius 1 to radius 3 (Figure 36D). Indeed, the distribution of errors in the simulated behavior was more precise for radius 3 than for radius 1 (Figure 38B). At first sight, this might look paradoxical, as a broader activity bump results in a narrower distribution of angle readouts from the bump (higher precision, Figure 38C). However, this can be explained by the fact that a larger bump is more efficient at averaging out noise fluctuations, so bump diffusion during the memory period is reduced and precision increased. The greater efficacy of a broad bump for

averaging out noise fluctuations and increase precision can be specifically assessed by running simulations for increasing amplitudes of external noise fluctuations (Figure 38D): the steeper dependency of angular error with noise amplitude for radius 1 than for radius 3 is consistent with narrower bumps being less effective at averaging out noise.

Memory errors found in behavior are not just a consequence of noise. In multi-item trials, errors occur because of the interference between the target and the NT item. Ring attractor models can produce the pattern of

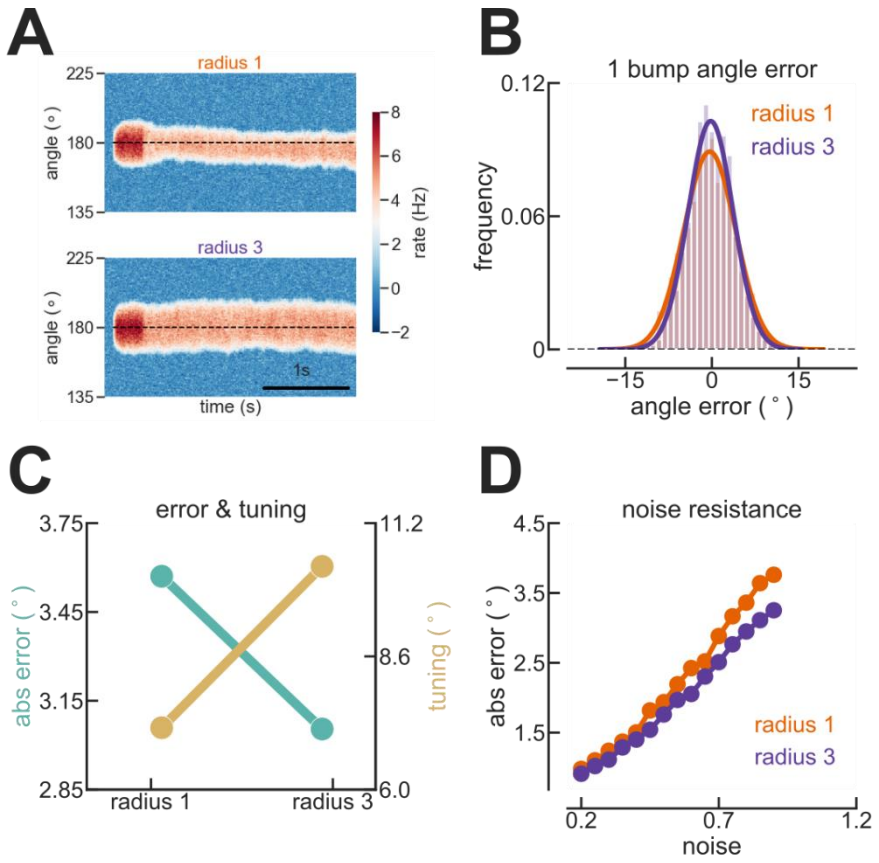

Figure 38 Model shows that the increased angular precision with radius can be explained by reduced angular tuning at larger eccentricities.

A) Sample single-item simulations of two network models, for stimuli presented at radius 1 (top), or radius 3 (bottom), differing just in connectivity parameters. **B)** End-of-delay error distribution (3275 simulations) was narrower for radius 3 than radius 1 (std of Gaussian fit: 4.48º radius 1, 3.88º radius 3), in contrast to broader activity bump in radius 3 during the delay. **C)** Opposite trends of angular error (left y-axis) and tuning width (right y-axis) with network connectivity broadening (x-axis) in model simulations. **D**) Steeper dependency of angular error with input noise amplitude for radius 1 than radius 3 shows that broader bumps (radius 3) are more robust to noise.

interference of simultaneous memories, where the effective connectivity that accounts for feedback excitation and feedback inhibition to excitatory neurons through the local network (Mexican-hat connectivity) can explain attraction for close-by locations and repulsion for longer inter-item distances (Almeida et al., 2015). I wondered how the widening of the connectivity for radius 3 affected this pattern of interference. To this end,

I ran 5,800 simulations with two simultaneous memory bumps initialized at random locations on the network, and I analyzed the pattern of memory errors at the end of a 1.5s delay period as a function of the distance between target and NT. For both radii, our simulations showed that memories are attracted for close-by locations and repelled for more distant locations (Figure 39A). However, changes in connectivity cause a displacement in these curves and I found a range of inter-item angular distances where interference was repulsive for radius 1 but attractive for radius 3 (Figure 39A-B).

The cross-over from repulsion to attraction upon connectivity widening can be understood on the basis of the effective Mexican-hat feedback connectivity onto excitatory neurons. Each item stored in the network contributes a feedback current, and the signed overlap between the feedback currents caused by the two items determines their attractive or repulsive interaction. For a fixed distance between two items, a widening of the Mexican-hat connectivity can make an overall negative overlap turn positive, and thus explain the qualitative change in interference effect. This is illustrated in our simulation data in Figure 39C-D. I extracted the average current inputs impinging onto excitatory neurons in the network (bump feedback currents) at the end of the delay for a single item in each of our networks (for radius 1 and 3) and plotted them twice, separated by 22º (corresponding to repulsion in radius 1 and attraction in radius 3, Fig. 5A), thus schematically assessing the effect in multi-item trials of this distance. For radius 1 networks, the positive overlap of the excitatory lobes of both bump feedback currents (pink shaded area) was smaller than the negative overlap between the inhibitory lobes of one bump feedback current with the excitatory lobe of the other (green shaded area), explaining the overall repulsive effect (Figure 39C). Instead, broader bump feedback currents in radius 3 networks led to larger positive than negative overlaps for the same inter-item distance, thus explaining the cross-over to an overall attractive effect (Figure 39D).

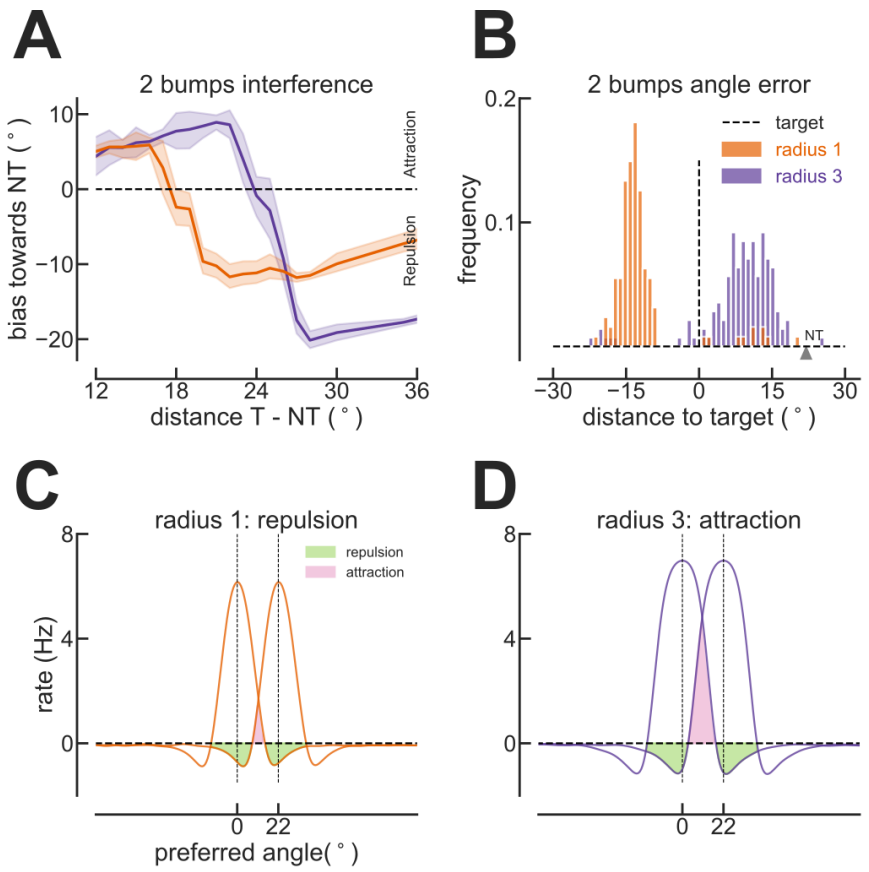

Figure 39 Reduced angular tuning at large eccentricities predicts interference effects in multi-item network simulations.

Interference effect in simulations with 2 simultaneous bumps for the two network models (radius 1/radius 3) of Figure 38. Interference effect (y-axis), measured as end-of-delay bump displacement in the direction of the other bump, versus distance between the two simultaneous bumps (x-axis) (n=15,000 simulations). For both networks, an attractive regime is followed by a repulsive regime that vanishes with distance (Almeida et al., 2015), but different connectivity widths lead to a window of stimulus distances (ca. 18º-24º) with qualitative different behaviors in the two networks: repulsion for radius 1 and attraction for radius 3. **B)** Distribution of errors in n=268 simulations with inter-item distance of 22° shows systematic attraction towards the NT bump in radius 3 networks (mean=8.6°) but systematic repulsion in radius 1 networks (mean=-11.7°). **C, D)** Explanation of cross-over from repulsive (**C**) to attractive (**D**) behavior as the network connectivity widens. Individual curves are delay-period current input to network neurons in corresponding single-item simulations and are superimposed here at the distance used for two-item simulations in **B**. Overlap of positive currents marked with pink shading, overlap of positive and negative currents marked with green shading. Smaller (larger) positive than negative overlap explains repulsion (attraction) in the simulations of panel B.

Network simulations suggest that the specific feature that distinguishes neural circuits responsible for memory maintenance from neural circuits responsible for sensory processing is a broadening of neuronal angular tuning for increasing eccentricities. In the visual cortex, neuronal receptive fields are known to increase size linearly with eccentricity (Burkhalter & Van Essen, 1986; Desimone & Schein, 1987; Felleman & Van Essen, 1987; Gattass et al., 1981; Toet & Levi, 1992), thus keeping a constant angular tuning. This invariant angular topography is consistent with the minimal eccentricity modulation of angular behavioral outputs in my delay-0 trials (Figure 36B and Figure 37A). Instead, memory requirements in delay-3 trials should engage a different network, with a topography characterized by tuning broadening with eccentricity. I tested this prediction by analyzing single-unit recordings in the PFC of 2 macaques performing an oculomotor delayed response task (*Methods-Electrophysiology*). Monkeys had to remember during a 3-second delay the location of a dot presented briefly at one of 8 fixed angles and one of 3 different eccentricities from the fixation point. Out of 516 single neurons recorded 153 showed significant angular tuning. I divided these 153 neurons according to their preferred radius (max firing rate) and calculated the angular tuning curve of each neuron at its preferred radius. The average tuning curves of

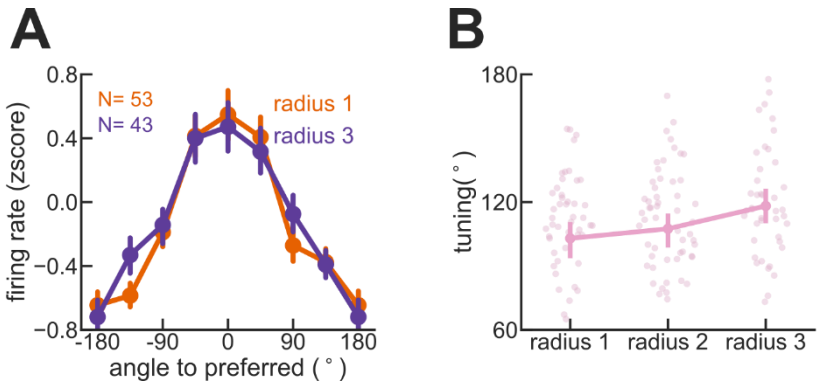

Figure 40 PFC single neurons show supralinear scaling of angular memory fields with preferred radius, as predicted by the computational model.

A) Average angular tuning curves for neurons with maximal responses in the smallest (radius 1, n=53) and largest (radius 3, n=43) radii tested reveal a small loss of angular tuning with preferred radius. The y axis represents z-scored firing rate, the x axis the angular distance of the stimulus to the neuron's preferred angle. **B)** Supra linear increase of PFC memory field size with preferred radius. The pink line shows broader tuning sorted by preferred eccentricity. The broader tuning as a function of radius (β=0.1307, t=2.420, ci=[0.024, 0.237] and p=0.017, n=153) shows a supralinear increase of memory fields with eccentricity in PFC.

neurons preferring each of the three radii showed broader tuning for the larger compared to the smaller radius (Figure 40A). Furthermore, a regression analysis showed that the widths of the neurons' tuning curves (circular standard deviation) increased with preferred radius (linear regression: n=153, β *intercept*=94.45, t=14.62, ci=[81.68, 107.21], p<0.001; β *eccentricity*=0.1307, t=2.420, ci=[0.024, 0.237], p=0.017, Figure 40B).

This confirms the prediction extracted from the computational model that motivated this analysis. In sum, electrophysiological recordings of PFC cortex concur with behavioral and modeling results in supporting different neural circuits for sensory and memory processing. While visual representations maintain a polar representation of the visual field, memory circuits are characterized by the known topography of visual areas (linear) could not account for the observed memory results, PFC topography (supra-linear) agrees both with behavioral and modeling evidence.

Radial dimension

In this section, I explored the topographical features of the vsWM circuit in the radial dimension. I analyzed the single-item trials of the vsWM task (Figure 35, *Methods-Paradigms and analysis*) and developed a radial implementation of the bump attractor model to explain the observed behavior. Therefore, instead of analyzing behavior in the angular dimension (as I did in *Results-Angular dimension*), I evaluated how the precision in the radial dimension depended on the eccentricity in both the no-delay and delay condition. The distribution of responses along the radial dimension (Figure 41A-B) clearly showed a decrease in precision for more eccentric trials for both the no-delay and the delay condition. Levene's test indicated differences in precision between radius ($t=20.64$, $p<0.001$ in the no delay condition and $t=32.94$, $p<0.001$ in the delay condition) and a mixed linear model with the individual standard deviation as dependent variable and eccentricity as independent variable confirmed this impairment with eccentricity in each condition (no delay: $\beta_{eccentricity}=0.046$, $z=9.93$, $ci=[0.037,0.055]$, $p<0.001$; delay: $\beta_{eccentricity}=0.080$, $z=5.86$, $ci=[0.083,0.107]$, $p<0.001$). As expected, the precision of the responses decreased with eccentricity, probably due to the fact the visual resolution in the fovea is higher than in the periphery.

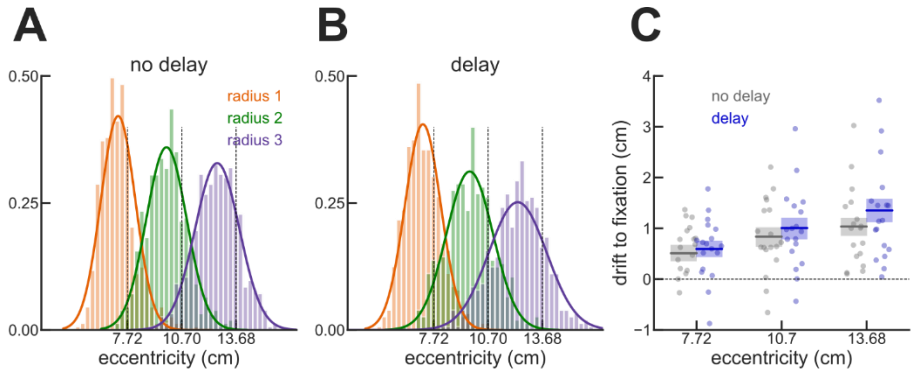

Figure 41 Behavioral compression of the visual space: eccentricity-delay interaction

A) Distribution of responses in the radial dimension for the no delay condition. Data show overall attraction towards fixation as well as a decrease in precision with eccentricity. **B)** Distribution of responses in the radial dimension for the delay condition. Both the decrease in precision with eccentricity and the attraction towards fixation are increased compared to the no delay condition. **C)** Quantification of the compression of the visual space for delay and eccentricity. The y axis represents the attraction toward the fixation point in the responses and the x axis the eccentricity of the targets. The small dots show the mean error in the radial dimension of each subject. I observed an overall attraction towards fixation and a significant interaction between delay and eccentricity (stats in main text). Results show an augmented compression of the visual space with delay.

Besides analyzing precision, I also checked for biases in the responses, as they could reveal topographical characteristics of the circuit. The distribution of responses showed a general mislocalization of the responses in the direction of the fixation point. For all the different conditions, I fitted a gaussian distribution and compared the center of the distribution with the actual target location. Figure 41A-C clearly show how this mislocalization was present for all the eccentricities in both the no delay and the delay condition. In Figure 41C, I quantified this attraction towards the fixation point as a function of both eccentricity and delay. The mixed linear model showed a significant interaction between eccentricity and delay (mixed linear model, N=18, n=3396, β *intercept*=-0.163, z=-0.82, ci=[-0.55, 0.225], p=0.412; β *eccentricity*=0.088, z=9.07, ci=[0.069,0.107], p<0.001; β *delay*=-0.078 , z=-1.554 , ci=[-0.177, 0.020], p=0.120; β *eccentricity x delay*=0.013, z=2.865, ci=[0.004, 0.022], p=0.004). I analyzed this interaction with independent mixed linear models for each delay condition. There was a significant effect of eccentricity for both the no-delay and delay condition (pvalues<0.001) but the slope for the eccentricity for the delay condition was higher than in the no-delay condition (no-delay: β *eccentricity*=0.088, ci=[0.071, 0.105]; delay: β *eccentricity*=0.128, ci=[0.107, 0.149]). Results show a compression of the visual space that gets stronger with eccentricity when the delay period is longer.

I hypothesized that a network with eccentricity-dependent changes in the connectivity parameters could explain behavioral results. I built an attractor network of excitatory and inhibitory rate-model neurons, recurrently connected through translationally-invariant Gaussian-like connectivity profiles (*Methods-Network model in the radial dimension*). This network was able to maintain the memory in the form of a stable bump during the delay period through reverberatory activity. Based on the hypothesis, I modeled the radial dimension by introducing changes in the connectivity parameters, leading to broader connectivity profiles for larger eccentricities (Figure 42A). This pattern of connectivity generated an attraction towards highly tuned neurons (fixation point) that evolved with delay (Figure 42B). The model reproduced the interaction between eccentricity and delay observed in behavior (Figure 41C), with a stepper mislocalization towards fixation with eccentricity for longer delay periods. I then evaluated the topographical structure of the network that produced this drift towards fixation. I observed that the key parameter of the model

was the exponent that determined the change of connectivity with eccentricity for both populations (b_x, *Methods-Network model in the radial dimension*). When this exponent was higher in the excitatory connections than in the inhibitory ones, I observed this mislocalization towards fixation that evolved with delay (Figure 42C). This attractive effect was stronger for further eccentricities when the difference between the exponents increased. Figure 42D-F show examples of simulations at different eccentricities to exemplify how this mislocalization towards fixation increases with eccentricity. Modeling analysis reproduced behavior and explained how the compression of the visual space evolved with delay due to a broadening of the connectivity profiles with eccentricity.

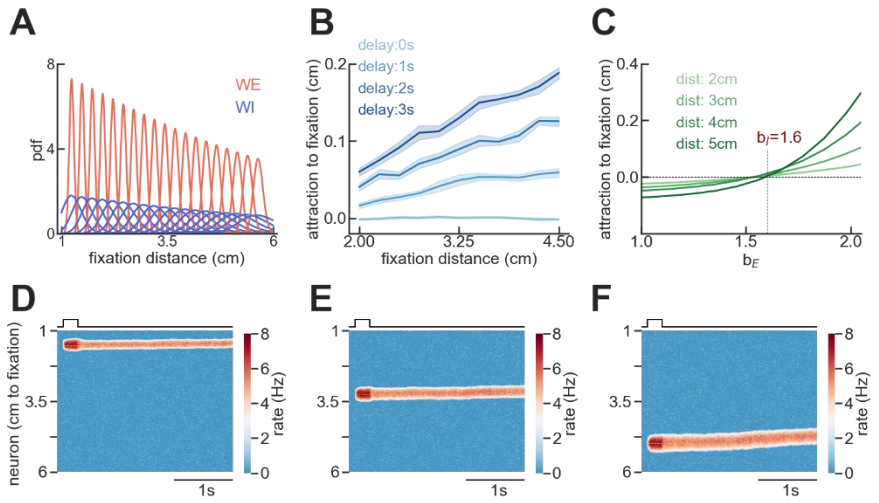

Figure 42 The bump attractor model predicts the eccentricity-delay interaction in the compression of the visual space.

A) Connectivity profiles for excitatory and inhibitory neurons coding for different eccentricities. Both profiles decay exponentially with eccentricity. **B)** Simulation results of the compression of the visual space with delay show an interaction between eccentricity and delay. Longer delays present increased attraction towards fixation with eccentricity (four different delay times -0s, 1s, 2s, 3s- and 9 different positions between 2cm and 4cm, n simulations=18000). The mechanistic explanation of this effect is illustrated in **C)** where I manipulated the exponential decay of the connectivity profile of excitatory neurons (simulations without noise to explore how the b_E parameter was responsible for the attractive or repulsive evolution of the effect. 4 different positions (2cm, 3cm, 4cm and 5cm; n simulations=100 with fixed delay of 1s). I observed that when this decay is stronger than the decay of the inhibitory neurons (fixed at 1.6), the compression of the visual space is present. **D**, **E** and **F)** Show examples of 3000ms simulations at increasing eccentricities. I clearly observe broader bumps with larger drift towards fixation for the most eccentric positions (**F**) compared to the less eccentric (**E**).

4.2. Distractor filtering in the working memory circuit

In this chapter, I addressed goals **(3)**, **(4)** and **(5)** through the investigation of how distracting information interferes with the WM content in behavioral, neuroimaging and electrophysiological data.

Although a "distractor" could be literally defined as "something that distracts", in this thesis distractors are defined as irrelevant information coexisting in the same task with the relevant information (targets). Among all the different domains where distractors can be manipulated (number, time, similarity, saliency…) I decided to explore *similarity,* by parametrically manipulating the angular distance between the target and the distractor; and time, by parametrically changing the time between the target and the distractor presentation -target-distractor onset asynchrony (TDOA)- and the order of appearance (target first or distractor first). I studied **(3)** the behavioral effects of this manipulations as well as their neural correlates in BOLD signal. These results are also related to the SRT by interpreting if the observed reconstructions using IEMs are compatible with this theory or if, alternatively, they support WM maintenance in frontal regions. Goal (3) is mainly explored in the sections *Distractor filtering: TDOA and order effects* and *Target and distractor reconstruction from BOLD signal*.

The manipulation of distance, order and TDOA allowed me to explore different mechanisms for distractor filtering using the bump attractor model **(4)**. I approached this on two fronts. For one, I simulated explicitly the utilized vsWM task and developed a control mechanism based on excitability control. Secondly, I analyzed the behavioral and the electrophysiological results of Qi et al. (2021) and developed a control mechanism based on neuromodulatory control. Goal **(4)** is mainly explored in the sections *Mechanistic explanation for distractor filtering* and *Distractor filtering under NB stimulation*.

The developed computational models to explain distractor filtering had some predictions that could be tested in neural activity **(5)**. This final goal was addressed by analyzing neural recordings in two different datasets (*Methods-Electrophysiology*, Dataset 2, and Dataset 3) and it is presented in the sections *Distractor filtering: electrophysiology* and *Distractor filtering under NB stimulation*.

Distractor filtering: TDOA and order effects

In this section, I developed a vsWM task where participants had to remember the spatial location of three stimuli during a delay period of 12s while ignoring distracting information (Figure 43). Distractors were parametrically modulated in time (order & TDOA) and similarity (angular distance target-distractor). Further details of the task can be found in *Methods-Paradigms and analysis.* Regarding the temporal domain, the task had 4 different conditions: *order 1-short TDOA*, *order 1-long TDOA*, *order 2-short TDOA and order 2-long TDOA*. Regarding the similarity domain, although the distance between targets and distractors was parametrically modulated, I will present the results in *close trials* (distractor in the same quadrant) and *far trials* (distractor in another quadrant).

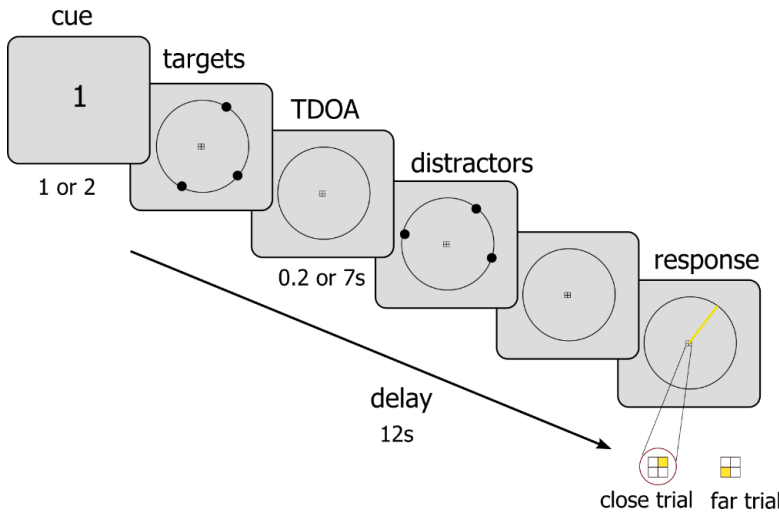

Figure 43 vsWM & distractor filtering task

Schematic view of a trial. Participants had to fixate at the center of the screen to start the trial. The fixation point consisted in 4 empty squares. Then, a cue indicating the order appeared (1: order 1, participants had to remember the first set of stimuli. 2: order 2, participants had to remember the second set of stimuli. TDOA was also manipulated (short TDOA=0.2s and long TODA=7s). In all four conditions (order 1 TDOA=0.2, order 1 TDOA=7, order 2 TDOA=0.2, order 2 TDOA=7) participants remembered the relevant information for 12s. The spatial distance between targets and distractors was also controlled. One distractor was located 10-20° away from the target in the same quadrant (close). Another was located 20-30° away from the target in the same quadrant (close). The last one was located 40-180° away from the target in a different quadrant (far). At the end of the delay period, one of the for squares of the fixation point turned yellow, indicating subjects had to report the location of the target that appeared in that quadrant. To report the exact location, subjects had to adjust a yellow bar that randomly appeared at one of the adjacent vertical or horizontal axis. More details in *Methods-Paradigms and analysis*.

I first investigated if distractor had an effect in the precision of the final reports depending on the temporal condition (order-TDOA). To do so, I calculated the angular error (target-response) in each condition and calculated both the mean and the standard deviation of the error distributions (Figure 44). Distractors were presented cw to the target in half of the trials and ccw to the target in the other half, so no systematic bias in the mean of the distributions was observed (paired ANOVA: n=26, independent variable: angular error, dependent variable: temporal conditions (order x TDOA), F=0.17, p=0.92). When checking the precision however, I observed a significant effect of the condition (Levene's test: W=4.66, p=0.003). Precisely, this difference came from the condition *order 1 – TDOA short,* which showed larger variance compared to the other

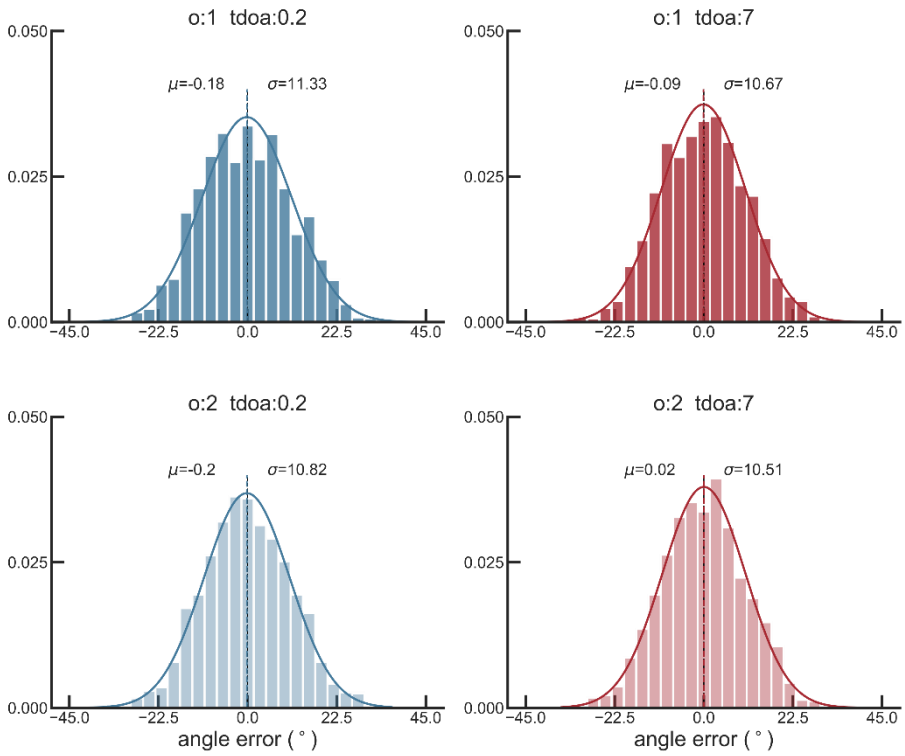

Figure 44 Distribution of errors per condition

Distribution of errors in all four temporal conditions (order 1-2 x TDOA 0.2-7s). Top row shows the distribution of errors in order 1 conditions (target presented before distractor) and the second row shows the distribution of errors in order 2 conditions (distractor presented before the target). Short TDOA conditions (0.2s) are represented in blue while long TDOA conditions (7s) are represented in red. I fitted a gaussian kernel for each distribution and calculated the mean and the standard deviation. *Order 1 – TDOA short* presented the larger variance.

conditions (Levene's test *order 1-TODA long*: W=6.61, p=0.01; *order 2-TODA short*: W=5.44, p=0.02; *order 2-TODA long*: W=12.95, p<0.001). Analyzing the absolute error with a mixed linear model provided analogous results (Figure 45), being again the order 1 – TDOA short the condition with larger error compared to the others (mixed linear model with random intercept per subject, N=27, n=6362, with the condition order 1 – TDOA 0.2 in the intercept: β order 1 – TDOA 7 =-0.513 , z=-2.503 , ci=[-0.914, -0.111] , p=0.012; β order 2 – TDOA 0.2 =-0.420 , z=-2.044 , ci=[-0.823, -0.017] , p=0.041; β order 2 – TDOA 7 =-0.718 , z=-3.526 , ci=[-1.118, -0.319] , p<0.001). The analysis of precision showed distractors had a differential impact when manipulating the temporal domain, being the *order 1 – TDOA short* condition the most distracting one.

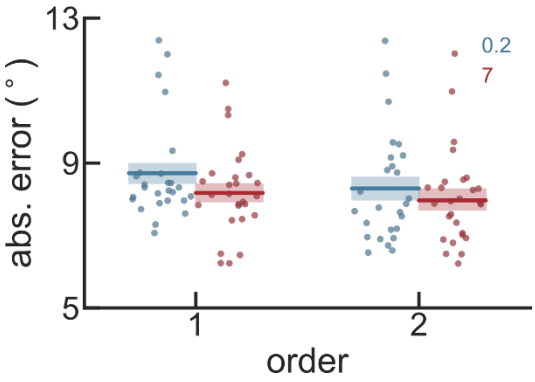

Figure 45 Absolute error per condition
Absolute errors in all four temporal conditions (order 1-2 x TDOA 0.2-7s). Each point is the mean absolute error of each subject (N=27). The box shows the population mean and the sem. Larger absolute errors were observed in the *Order 1 – TDOA short.*

Then, I investigated if distractors interfered with the targets differently depending on the spatial distance separating them. At this point, it is important to remind that this dataset is a combination of two, as the initial dataset was obtained by running the paradigm in the laboratory facility for psychophysical experiments (n=4475, N=21) and the other dataset was posteriorly obtained from running the paradigm inside the scanner (n=1887, N=6). In the first dataset, I parametrically modulated the separation in three different regimes (10-20° in the same quadrant, 20-30° in the same quadrant and 40-180° in a different quadrant). I then calculated the direction of the interference (attractive: responses are biased in the direction of the distractor, repulsive: responses are biased in the opposite direction of the distractor) in each regime (Figure 46) and observed attraction for the regimes of distraction in the same quadrant

(close trials: 10-20° and 20-30°) and repulsion for the regime of distraction in a different quadrant. A mixed linear model with the attraction towards the distractor as dependent variable and the three different regimes as independent variable (10-20° regime in the intercept, N=21, n=4475) showed a small increase of attraction in the 20-30° regime (β *20-30°*=0.407 , z=1.991, ci=[0.006, 0.809] , p=0.05) and a strong difference with the 40-180° regime (β *40-180°*=-4.099 , z=-21.345, ci=[-4.475, -3.722] , p<0.001). Based on these results, when I posteriorly ran the same task inside the scanner, the intermediate regime was eliminated used (scanner: 10-20° in the same quadrant and 40-180° in a different quadrant). In sum, distractors interfered in an attractive way when the distractor was located close to the target and in a repulsive way when they were located far away.

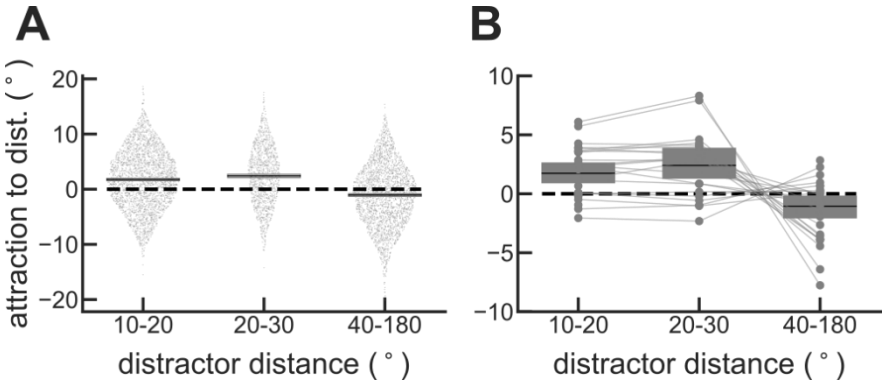

Figure 46 Attractive interference for close distractors and repulsive for distant ones
A) Distribution of interference towards the distractor in 4475 trials. Box showing the mean and the 95% ci. **B)** Same as A, but each point is the mean interference towards the distractor per subject. Attractive interference is observed for 10-20° and 20-30° distances ("close trials": inside the same quadrant) and repulsive interference is observed for 40-180° distances ("far trials": outside the quadrant).

When I ran a mixed model with order, TDOA and distance as regressors, I observed significant interactions between distance-order and distance-TDOA (mixed linear model with random intercept per subject, N=27, n=3920, β *intercept*= 3.627, z=8.474, ci=[2.788, 4.466], p<0.001; β distance= -5.782, z=-8.945, ci=[-7.049, -4.515], p<0.001; β order=-0.882, z=-3.477, ci=[-1.379, -0.385], p=0.001; β *TDOA*=-0.222, z=-2.758, ci=[-0.379, -0.064], p=0.006; β *distancexorder*=1.221, z=2.988, ci=[0.420, 2.021], p=0.003); β *distancexTDOA*=0.447, z=3.452, ci=[0.193, 0.701], p=0.001); β *orderxTDOA*=0.081, z=1.608, ci=[-0.018, 0.181], p=0.108); β

distancexorderxTDOA=-0.127, z=-1.559, ci=[-0.288, 0.033], p=0.119), indicating I should differentiate between close and far trials to correctly evaluate the different temporal conditions (*order 1-short TDOA*, *order 1-long TDOA*, *order 2-short TDOA and order 2-long TDOA*). Figure 47A shows interference towards the distractor for close trials (distracted in the same quadrant) while Figure 47B does it for far trials (distracted in a different quadrant). For close trials, I observed an effect of order, where presenting the distractor after the target (order 1) had larger attractive interference than presenting it before target presentation (order 2). I also effect an effect of TDOA, where short time (0.2s) between the target and the distractor had larger attractive effects than longer times (7s). A linear mixed model showed a trend towards and interaction order-TDOA, where the effect of TDOA is more noticeable in the order 1 condition (mixed linear model with random intercept per subject, N=27, n=3920, β *intercept*= 3.528, z=6.285, ci=[2.428, 4.628], p<0.001; β *order*=-0.849, z=-3.725, ci=[-1.295, -0.402], p<0.001; β *TDOA*=-0.216, z=-2.985, ci=[-0.357 -0.074], p=0.003; β *orderxTDOA*=0.083, z=1.831, ci=[-0.006, 0.173], p=0.067). For far trials, I observed and effect of TDOA, where short time (0.2s) between the target and the distractor had larger repulsive effects than longer times (7s). I did not observe any difference in order nor in the interaction order-TDOA (mixed linear model with random intercept per subject, N=27, n=2442, β *intercept*= -2.357, z=-3.456, ci=[-3.694, -1.020],

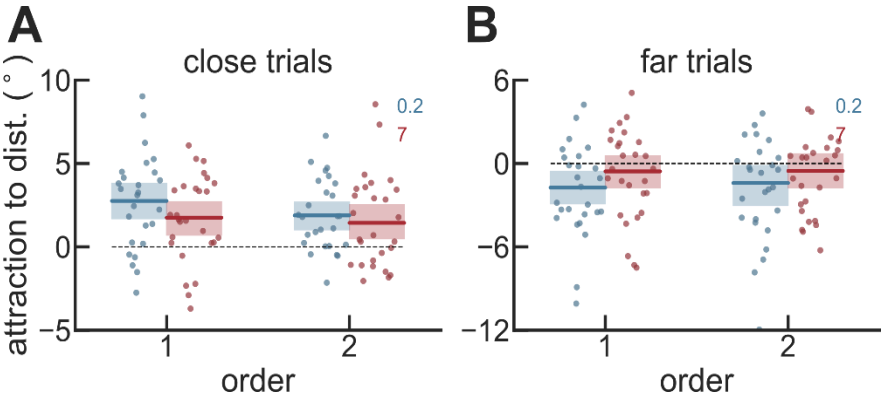

Figure 47 Interference effects towards distractors for distance, order and TDOA

A) Overall attractive interference toward the distractor for close trials. The box shows the population mean and the 95% ci. Each dot is the mean of a subject (n=27). Effect of order (larger attraction in order 1) and effect of TDOA (larger attraction for short TDOAs). Trend in the interaction order-TDOA (p=0.067) **B)** Overall repulsive interference against the distractor for far trials. Effect of TDOA, with larger repulsion for short TDOAs.

p=0.001; β *order*=0.359, z=1.181, ci=[-0.237, 0.955], p=0.238; β *TDOA*=0.204, z=2.118, ci=[0.015, 0.393], p=0.034; β *orderxTDOA*=-0.048, z=-0.795, ci=[-0.168, 0.071], p=0.427).

Figure 48 shows the effects of order and TDOA in all datasets, connecting the mean of each subject. Figure 48A-B shows the average of the datasets (same as in Figure 47), Figure 48C-D shows the results in the dataset obtained in the laboratory facility for psychophysical experiments (n=4475, N=21), and Figure 48E-F shows the results in the dataset obtained inside the scanner (n=1887, N=6). In both datasets, analyzing the results differently for close and far trials was justified after running a mixed linear model with distance, TDOA and order as regressors and observing significant effects in the interactions of distance with either order or TD0A (mixed model laboratory facility: N=21, n=4475, β *TDOAxdistance*= 0.321, z=2.034, ci=[0.012, 0.630], p=0.042; β *orderxdistance*= 0.681, z=1.358, ci=[-0.302, 1.664], p=0.174; mixed model scanner: N=6, n=1887, β *TDOAxdistance*=0.799, z=3.544, ci=[0.357, 1.241], p<0.001; β *orderxdistance*= 2.027, z=2.908, ci=[0.661, 3.393], p=0.004).

In the dataset obtained in the laboratory facility, I observed attractive interference and significant effects of order and TDOA, but no interaction for the close trials (mixed linear model with random intercept per subject, N=21, n=2981, β *intercept*= 3.972, z=6.079, ci=[2.691, 5.252], p<0.001; β order=-0.911, z=-3.534, ci=[-1.417, -0.406], p<0.001; β *TDOA*=-0.168, z=-2.076, ci=[-0.327, -0.009], p=0.038; β *orderxTDOA* =0.052, z=1.024, ci=[-0.048, 0.153], p=0.306). For the far trials, I observed repulsive interference and significant effect of TDOA once the no-significant interaction TDOA-order (β orderxTDOA =0.074, z=0.951, ci=[-0.078, 0.227], p=0.341) is removed from the model (mixed linear model with random intercept per subject, N=21, n=1494, β *intercept*= -2.467, z=-3.465, ci=[-3.862, -1.071], p=0.001; β order=-0.019, z=-0.072, ci=[-0.537, 0.499], p=0.942; β *TDOA*=0.182, z=4.637, ci=[0.105, 0.259], p<0.001). In the dataset obtained in the scanner, I observed attractive interference and significant effects of order and a trend for the interaction order-TDOA for the close trials (mixed linear model with random intercept per subject, N=6, n=939, β *intercept*= 2.028, z=2.158, ci=[0.186, 3.870], p=0.031; β *order*=-0.664, z=-1.375, ci=[-1.610, 0.283], p=0.169; β *TDOA*=-0.387, z=-2.453, ci=[-0.696, -0.078], p=0.014; β *orderxTDOA* =0.192, z=1.946, ci=[-0.001, 0.385], p=0.052). For the far trials, I observed significant effects of order, TDOA and the

interaction (mixed linear model with random intercept per subject, N=6, n=948, β *intercept*= -1.928, z=-1.981, ci=[-3.834, -0.021], p=0.048; β order=1.362, z=2.829, ci=[0.419, 2.306], p=0.005; β *TDOA*=0.410, z=2.653, ci=[0.107, 0.713], p=0.008); β *orderxTDOA*=-0.238, z=-2.438, ci=[-0.430, -0.047], p=0.015).

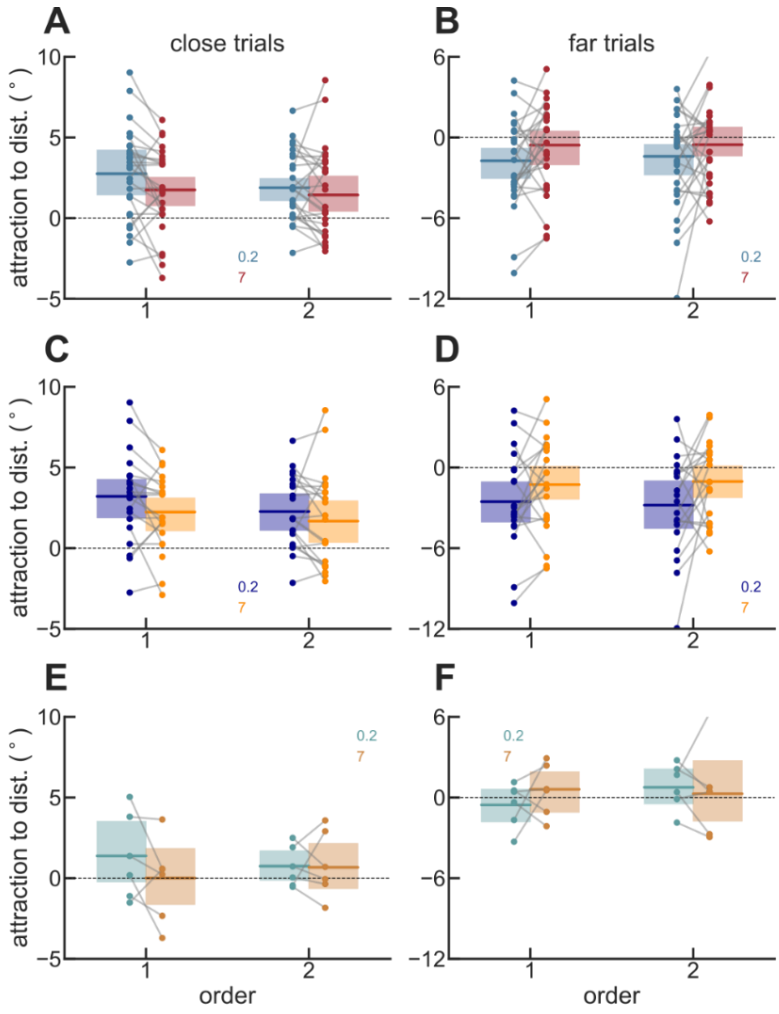

Figure 48 Behavioral effects in all the experimental settings (combined, laboratory and scanner)
A) Behavioral results in close trials. The plot is the combination of two datasets (dataset 1: obtained in the laboratory facility for psychophysical experiments, dataset 2: obtained inside the scanner). The box shows the mean interference and the 95% ci. Each point is the mean interference of each subject. The grey line connects the behavior of the same subject in each order condition. **B)** Behavioral results in the far trials (combination of both datasets). **C)** Behavioral results in the close trials for the dataset 1 (laboratory facility for psychophysical experiments). **D)** Behavioral results in the far trials for the dataset 1. **E)** Behavioral results in the close trials for the dataset 2 (inside the scanner). **F)** Behavioral results in the far trials for the dataset 2.

Target and distractor reconstruction from BOLD signal

In this section, I used an inverted encoding model (IEM) to generate model-based reconstructions of both the remembered (target) and ignored (distractor) angles from the BOLD signal patters in visual, parietal, and frontal areas. The visual area was defined by combining and atlas with retinotopy procedures while parietal and frontal areas were defined by combining and atlas with a localizer task (detailed description of the model and ROIs definition in *Methods-MRI*). I decided to evaluate the strength of the reconstructions in these three regions based on previous studies (Ester et al., 2015; Rademaker et al., 2019; Serences, 2016; Sprague et al., 2014).

Target reconstruction

I first focused on reconstructing the remembered angle during the delay period of each of the four temporal conditions (*order 1-short TDOA*, *order 1-long TDOA*, *order 2-short TDOA and order 2-long TDOA*). Figure 49A show a schematic view of all the temporal conditions of the task (Figure

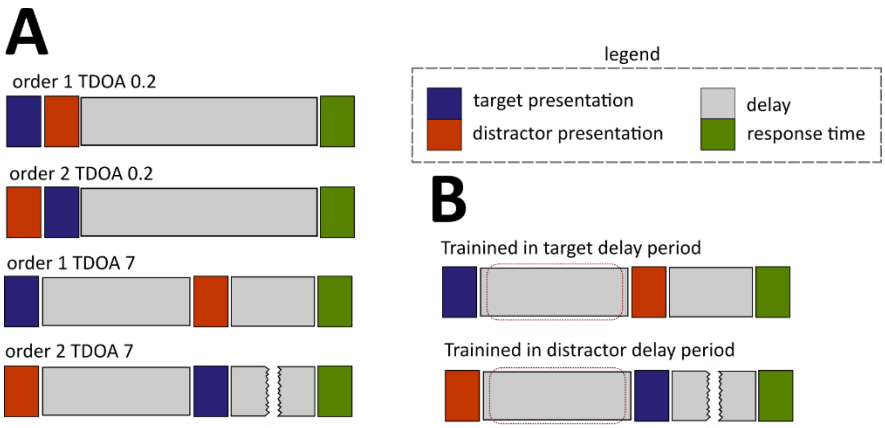

Figure 49 Schematic view of the vsWM task periods

A) Schematic view of all four temporal conditions of the vsWM task. From top to bottom: *order 1-TDOA=0.2s; order 2-TDOA=0.2s; order 1-TDOA=7s and order2-TDOA=7s*). In all the conditions, the delay period after the target presentation was of 12s. To avoid confusion, the second delay of the condition *order 2-TDOA=7s* appears broken, as its duration was 12s. **B)** Training protocols. When reconstructing the target, the IEM was trained in the first delay period of the condition *order 1-TDOA=7s (target delay period)* and it was tested in all the other conditions (cross-validating) When reconstructing the distractor, the IEM was trained in the first delay period of the condition *order 2-TDOA=7s (distractor delay period)* and it was tested in all the other conditions (cross-validating).

43). As fMRI has a low temporal resolution compared to other neuroimaging techniques (electrophysiology, EEG), I could not evaluate the reconstruction of the remembered angle in the TDOA of the TDOA-short conditions (Figure 49A, top two rows). On the other hand, a long TDOA of 7s was used to evaluate the reconstruction in this period (Figure 49A, bottom two rows). I was interested in evaluating the reconstructions under two different training protocols (Figure 49B): training in the *target delay period* (first delay of the condition *order 1-long TDOA*) and training in the *distractor delay period* (first delay of the condition *order 2-long TDOA*). I decided to use these periods for training as they were the only ones where either the target or the distractor were presented with no concomitant information. By doing so, I was able to evaluate the existence of a shared code for the target and the distractor in different regions of the cortex.

When training in the *target delay period* I could systematically reconstruct the remembered angle during the delay period in all four temporal conditions in visual, parietal, and frontal areas. Figure 50A shows the average reconstruction strength (decoding) during the delay period in each condition. For the condition *order 1-TDOA=0.2s*, I averaged the TRs comprising the delay period following the distractor presentation until the response time (Figure 49A, top row, delay). For the condition *order 2-TDOA=0.2s*, I averaged the TRs comprising the delay period following the target presentation until the response time (Figure 49A, second row, delay). For the condition *order 1-TDOA=7s*, I averaged the TRs comprising the delay period following the target presentation until the distractor presentation (Figure 49A, third row, first delay). For the condition *order 2-TDOA=7s*, I averaged the TRs comprising the delay period following the target presentation until the response presentation (Figure 49A, bottom row, second delay). Decoding was evaluated individually for each subject and its value is the difference between the reconstruction of the remembered angles with the reconstruction of no information (*Methods-MRI*). Significance was tested independently for each subject using permutation tests.

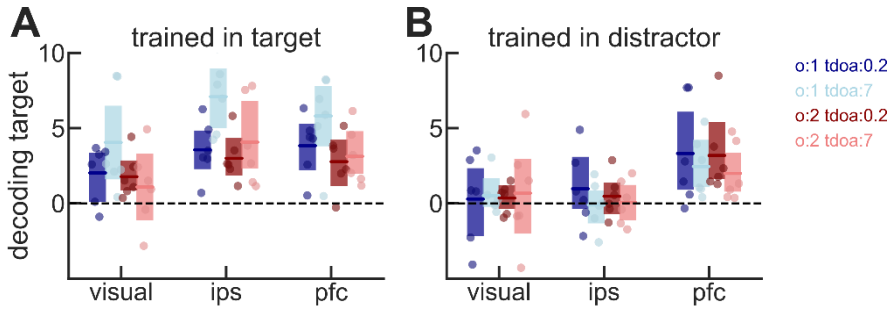

Figure 50 Target decoding based on training paradigm

A) Decoding of the target in all four temporal conditions in visual, parietal, and frontal areas when training the IEM in the *target delay period*. The box shows the mean decoding and the 95% ci. Each dot is the mean decoding of each subject. When training in the *target delay* period, the target could systematically be reconstructed in all the regions and all the conditions. The condition where the target was presented with no concomitant information (*order 1-TDOA=7*) presented higher decoding values. B) Decoding of the target in all four temporal conditions when training the IEM in the *distractor delay period*. Just frontal areas allow training in the *distractor delay period* to systematically reconstruct the target.

In visual (Figure 50A, left), when decoding the target training in the *target delay period*, I observed a significant reconstruction of the angle in the delay period of three out of the four conditions (*order 1-short TDOA* (t=3.415, p=0.002), *order 1-long TDOA* (t=3.822, p=0.002), *order 2-short TDOA* (t=3.91, p<0.001), *order 2-long TDOA* (t=1.637, p=0.115). In parietal (Figure 50A, middle), when decoding the target training in the *target delay period*, I observed a significant reconstruction of the angle in the delay period of all four conditions (*order 1-short TDOA* (t=4.695, p<0.001), *order 1-long TDOA* (t=9.422, p=<0.001), *order 2-short TDOA* (t=5.382, p<0.001), *order 2-long TDOA* (t=5.218, p<0.001). In frontal (Figure 50A, right), when decoding the target training in the *target delay period*, I observed significant reconstruction of the angle in the delay period of all four conditions (*order 1-short TDOA* (t=5.078, p<0.001), *order 1-long TDOA* (t=6.170, p=<0.001), *order 2-short TDOA* (t=4.864, p<0.001), *order 2-long TDOA* (t=5.029, p<0.001).

When training in the *distractor delay period* I could just systematically reconstruct the remembered angle during the delay period in frontal areas, and not in visual and parietal (Figure 50B). In visual (Figure 50B, left), when decoding the target training in the *distractor delay period*, I did not observe significant reconstruction of the angle in the delay period of any of all four conditions (*order 1-short TDOA* (t=0.459, p=0.650), *order 1-long*

TDOA (t=1.275, p=0.228), *order 2-short TDOA* (t=0.882, p=0.387), *order 2-long TDOA* (t=0.767, p=0.451). In parietal (Figure 50B, middle), when decoding the target training in the *distractor delay period*, I did not observe significant reconstruction of the angle in the delay period of any of all four conditions *order 1-short TDOA* (t=1.432, p=0.166), *order 1-long TDOA* (t=-0.19, p=0.852), *order 2-short TDOA* (t=0.751, p=0.460), *order 2-long TDOA* (t=0.093, p=0.926). In frontal (Figure 50B, right), when decoding the target training in the *distractor delay period*, I observed significant reconstruction of the angle in the delay period of all four conditions (*order 1-short TDOA* (t=2.824, p=0.009), *order 1-long TDOA* (t=2.846, p=0.016), *order 2-short TDOA* (t=3.889, p<0.001), *order 2-long TDOA* (t=3.04, p=0.005).

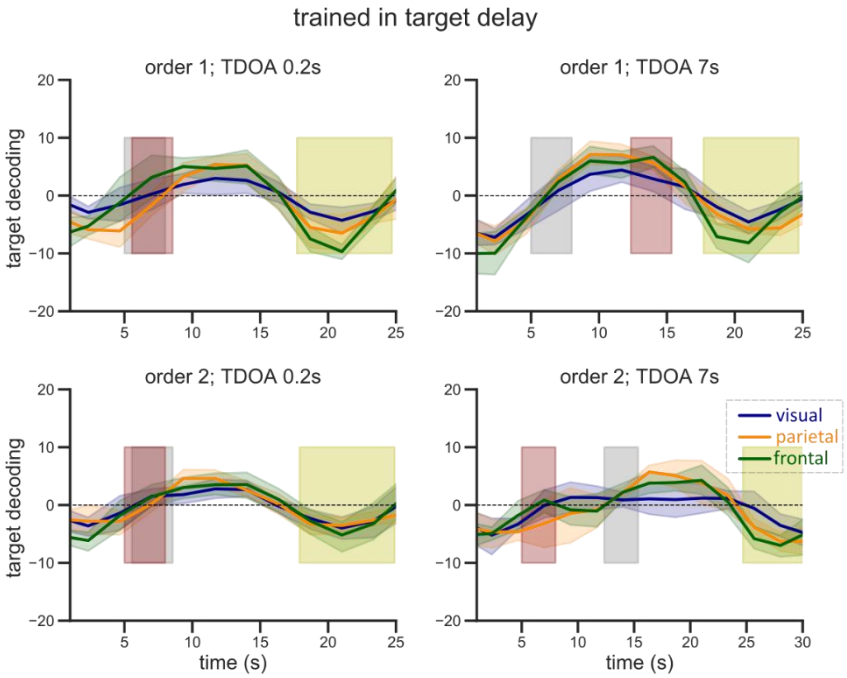

Figure 51 Time evolution of target decoding when training in the target delay period
Target decoding in all four temporal conditions when training in the *target delay period*. Boxes represent the different events: grey box represent the presentation of the target; the red box represents the presentation of the distractor and the yellow box the response time. All the events consider a 4s delay due to the hemodynamic response and present a window of 3s to correct for individual variability. The three lines show the mean decoding of the six participants with a 95% ci in the three ROIs (visual, parietal, and frontal areas).

Despite the low temporal resolution of fMRI, I could still analyze the temporal evolution of the decoding in each condition. I used a fixed TR of 2.335s, meaning each recording is spaced by this period. Figure 51 shows the decoding of the target when training in the *target delay period*. For all conditions, the decoding is higher during the first interval of the delay, which is consistent with the training protocol (first delay of *order 1 – long TDOA*). The figure shows the mean decoding of all six participants in the three regions of interest, with a 95% ci. As the measure of decoding is the difference between the reconstruction of the remembered angles with the reconstruction of no information (*Methods-MRI*), just positive significant positive values indicate a positive reconstruction. Results show the remembered angles can be reconstructed in all three regions, peaking at the beginning of the delay period and slowly decaying as the delay evolves. Extended results regarding the percentage of significant TRs during the delay period in each subject can be found in Table 1-Table 4. Results of Figure 51 are extended in Figure 52 and Figure 53. Figure 52 shows the time evolution of the decoding in each participant for the order 1 condition. On the other hand, Figure 53 shows it for the order 2 condition. The size of the dots represent the number of subjects that presented significant decoding in that specific TR, using a permutation test. For both the order 1 and order 2 conditions, subjects presented similar patterns, all presenting the peak of decoding at the beginning of the delay period.

training target	*visual*	*parietal*	*frontal*
s1	100%	100%	100%
s2	75%	75%	75%
s3	50%	50%	75%
s4	75%	100%	100%
s5	25%	50%	25%
s6	75%	75%	75%
training distractor	***visual***	***parietal***	***frontal***
s1	0%	50%	50%
s2	100%	50%	75%
s3	0%	75%	50%
s4	25%	25%	75%
s5	100%	50%	75%
s6	75%	75%	75%

Table 1 Decoding target order 1 - short TDOA

Decoding of target information training in the target and the distractor delay period in the condition order 1 -short TDOA. Percentage of significant TRs in each subject during the delay period using a permutation test.

training target	*visual*	*parietal*	*frontal*
s1	50%	100%	100%
s2	100%	100%	100%
s3	100%	100%	100%
s4	100%	100%	100%
s5	100%	100%	50%
s6	50%	100%	100%
training distractor	*visual*	*parietal*	*frontal*
s1	0%	100%	50%
s2	50%	50%	50%
s3	50%	0%	100%
s4	0%	50%	100%
s5	50%	0%	50%
s6	100%	50%	50%

Table 2 Decoding target order 1 - long TDOA

Decoding of target information training in the target and the distractor delay period in the condition order 1 -long TDOA. Percentage of significant TRs in each subject during the delay period using a permutation test.

training target	*visual*	*parietal*	*frontal*
s1	100%	100%	75%
s2	75%	50%	100%
s3	75%	75%	75%
s4	100%	100%	100%
s5	25%	75%	50%
s6	50%	100%	75%
training distractor	*visual*	*parietal*	*frontal*
s1	25%	50%	50%
s2	25%	50%	50%
s3	25%	75%	50%
s4	75%	0%	100%
s5	0%	50%	75%
s6	75%	50%	75%

Table 3 Decoding target order 2 -short TDOA

Decoding of target information training in the target and the distractor delay period in the condition order 2 -short TDOA. Percentage of significant TRs in each subject during the delay period using a permutation test.

training target	*visual*	*parietal*	*frontal*
s1	25%	50%	75%
s2	50%	100%	100%
s3	100%	50%	50%
s4	75%	100%	100%
s5	50%	100%	75%
s6	100%	100%	100%
training distractor	*visual*	*parietal*	*frontal*
s1	100%	50%	50%
s2	50%	0%	100%
s3	25%	50%	50%
s4	0%	50%	25%
s5	50%	0%	50%
s6	100%	50%	75%

Table 4 Decoding target order 2 -long TDOA

Decoding of target information training in the target and the distractor delay period in the condition order 2 -long TDOA. Percentage of significant TRs in each subject during the delay period using a permutation test.

The use of IEM allowed me to evaluate whether the pattern of BOLD signal associated with WM maintenance was consistent with a dynamic coding of WM or, alternatively, with a stable coding (Barbosa, 2017; Spaak et al., 2017). While a dynamic coding of WM would suggest different BOLD patterns from trial to trial and across temporal conditions, a stable coding would suggest a shared pattern of BOLD signal associated with memory maintenance. As the IEM was trained in a specific period of just one of the temporal conditions (first delay or *order 1-long TDOA*), the results presented in Figure 50A and Figure 51 suggest a stable coding for WM maintenance, as the pattern of BOLD signal of one condition successfully decode the remembered angles in other temporal conditions. Besides studying the stability of WM content, I could also inspect whether the patterns of BOLD signal during the delay could be used to reconstruct the remembered angles during other intervals of the task. Figure 52 and Figure 53 showed that decoding was mostly restricted to the delay period, so training in the BOLD signal of the delay period was not useful to reconstruct the remembered angles during neither the stimulus presentation nor the response time.

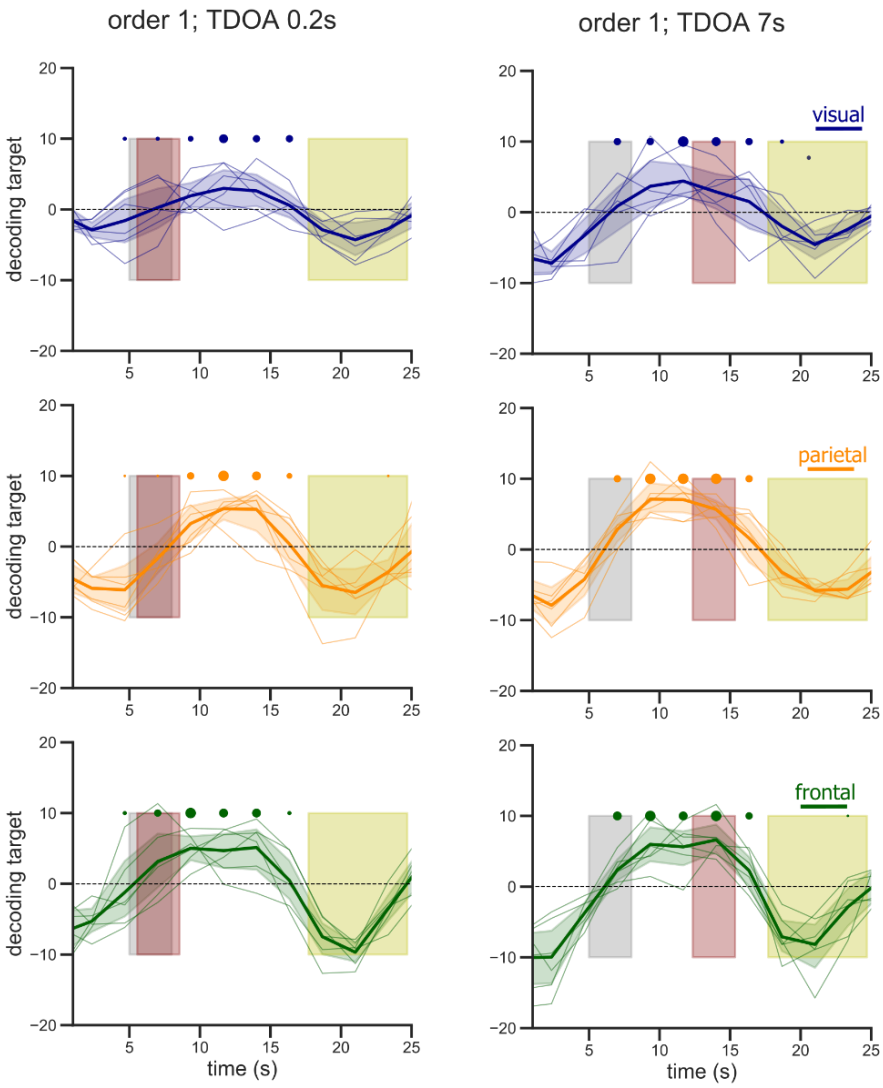

Figure 52 Individual target decoding in order 1 when training in the target delay period
Target decoding in order 1 conditions. Small linewidth lines show the individual decoding, and the large linewidth line shows the mean of them with 95% ci. Boxes represent the different events: grey box represent the presentation of the target; the red box represents the presentation of the distractor and the yellow box the response time. The dots on top of the TRs shows the number of subjects with significant decoding. The size of the dot linearly increases with the number of subjects with significant decoding (permutation test, α=0.05).

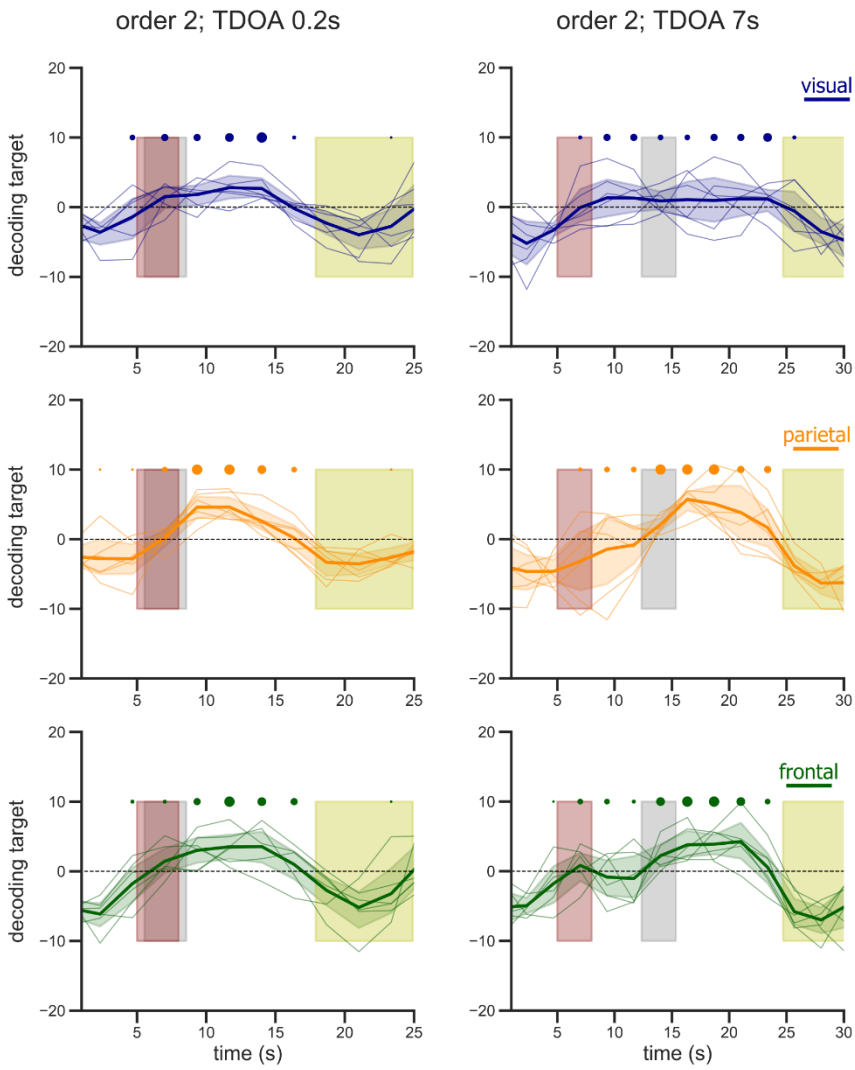

Figure 53 Individual target decoding in order 2 when training in the target delay period
Target decoding in order 2 conditions. Small linewidth lines show the individual decoding, and the large linewidth line shows the mean of them with 95% ci. Boxes represent the different events: grey box represent the presentation of the target; the red box represents the presentation of the distractor and the yellow box the response time. The dots on top of the TRs shows the number of subjects with significant decoding. The size of the dot linearly increases with the number of subjects with significant decoding (permutation test, α=0.05).

Then, I also analyzed the time evolution of target decoding when training in the *distractor delay period* (first delay of *order 2 – long TDOA*). Figure 50B showed just frontal regions could decode the target under this training protocol, especially in the short TDOAs conditions. Figure 54 shows the mean decoding of all six participants in the three regions of interest, with a 95% ci. While visual and parietal did not present significant decoding during the delay period following the target presentation, frontal regions did it, again showing the better reconstruction in the initial TRs of the delay period.

I observed significant decoding for the target before it is presented in the condition *order2-long TDOA*. As the training-testing protocol controlled for any possible contamination even within the same running session (I

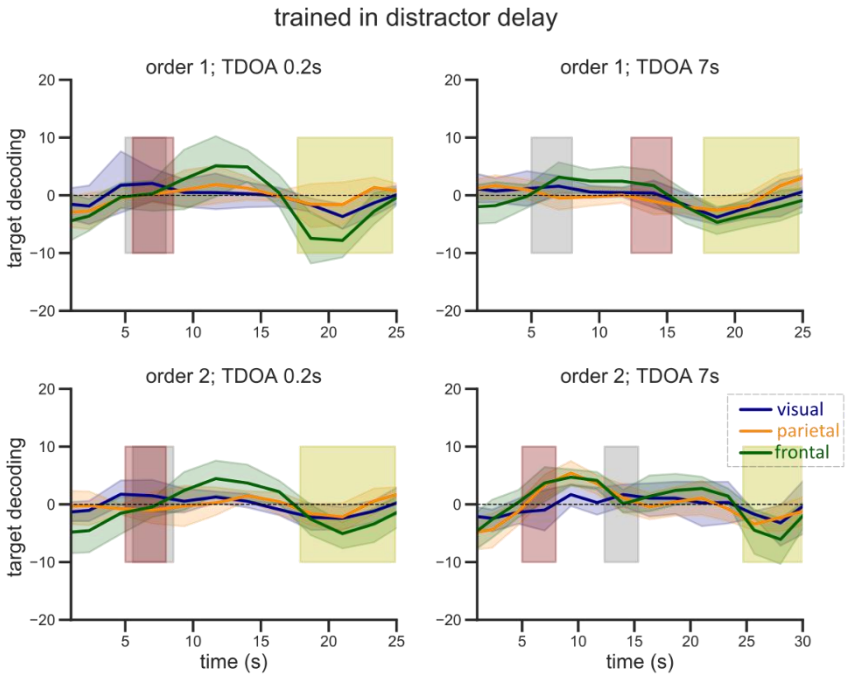

Figure 54 Time evolution of target decoding when training in the distractor delay period
Target decoding in all four temporal conditions when training in the *distractor delay period*. Boxes represent the different events: grey box represent the presentation of the target; the red box represents the presentation of the distractor and the yellow box the response time. All the events consider a 4s delay due to the hemodynamic response and present a window of 3s to correct for individual variability. The three lines show the mean decoding of the six participants with a 95% ci in the three ROIs (visual, parietal, and frontal areas).

tested in different scan sessions than the trained ones), I interpreted this signal as an anticipation of the position of the target combined with a low spatial resolution of the IEM. Anticipating the approximate position of the target stimulus that I would later try to decode was easy in this specific trial type, as stimuli were always presented in three different quadrants and, after the presentation of the distractors, the remaining quadrant without stimulus after the presentation of the distractors would always present one of the targets. Although it would only be the one the participants had to report one third of the times, the target presented alone is the one I try to decode in this analysis (to avoid confounding the decoding of the target with the decoding of the distractor).

Both the results of the reconstruction of the target during the delay period in each condition (Figure 50) and the time evolution of this decoding under different training protocols (Figure 51 and Figure 54) illustrate a clear difference between frontal areas and visual or parietal. While reconstructing the remember angle in frontal regions was independent of the training protocol, in visual and parietal regions, the reconstruction was only possible when training in the *target delay period.* Figure 55 shows the mean decoding of the target during the delay period following the target presentation depending on the training protocol (visual: $t=3.25$, $p=0.001$; parietal: $t=7.11$, $p<0.001$; frontal: $t=1.42$, $p=0.16$). As opposed to Figure 50, it averages the TRs of the different temporal conditions because, as observed in the time evolution of the decoding, barely any difference is observed in the dynamics after the target presentation.

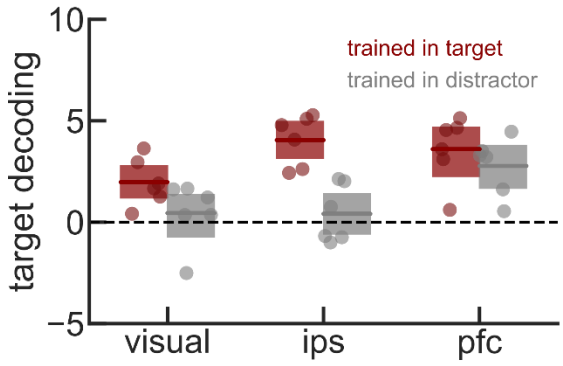

Figure 55 Frontal areas can decode the target independently of the training protocol.
Decoding of the target in visual, parietal, and frontal areas when training the IEM either in the *target delay period* or in the *distractor delay period.* The box shows the mean decoding of all six subjects and the 95% ci. Each dot is the mean decoding of each subject during the delay period following the target presentation.

Distractor reconstruction

In the previous sub-section, I presented the results of trying to reconstruct the remembered relevant information for the task (target). In this section, I show the results of reconstructing the irrelevant information (distractors). I present both the time evolution of the reconstructions and the mean decoding in the delay periods following the distractor presentation. In this last scenario, I take different times depending on the condition: for *order 1-TDOA=0.2s*, I averaged the TRs comprising the delay period following the distractor presentation until the response time (Figure 49A, top row, delay). For the condition *order 2-TDOA=0.2s*, I averaged the TRs comprising the delay period following the target presentation until the response time (Figure 49A, second row, delay). For the condition *order 1-TDOA=7s*, I averaged the TRs comprising the delay

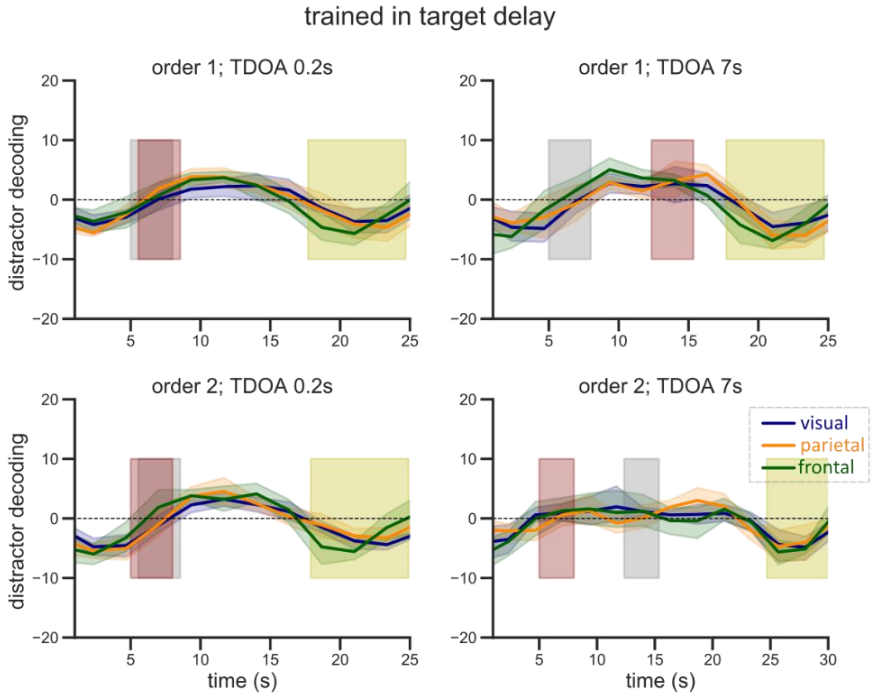

Figure 56 Time evolution of distractor decoding when training in the target delay period
Distractor decoding in all four temporal conditions when training in the *target delay period*. Boxes represent the different events: grey box represent the presentation of the target; the red box represents the presentation of the distractor and the yellow box the response time. All the events consider a 4s delay due to the hemodynamic response and present a window of 3s to correct for individual variability. The three lines show the mean decoding of the six participants with a 95% ci in the three ROIs (visual, parietal, and frontal areas).

period following the distractor presentation until the response presentation (Figure 49A, third row, second delay). For the condition *order 2-TDOA=7s*, I averaged the TRs comprising the delay period following the distractor presentation until the target presentation (Figure 49A, bottom row, first delay). Decoding was evaluated individually for each subject and its value was the difference between the reconstruction of the irrelevant angles (distractors) with the reconstruction of no information (*Methods-MRI*). Significance was tested independently for each subject using permutation tests.

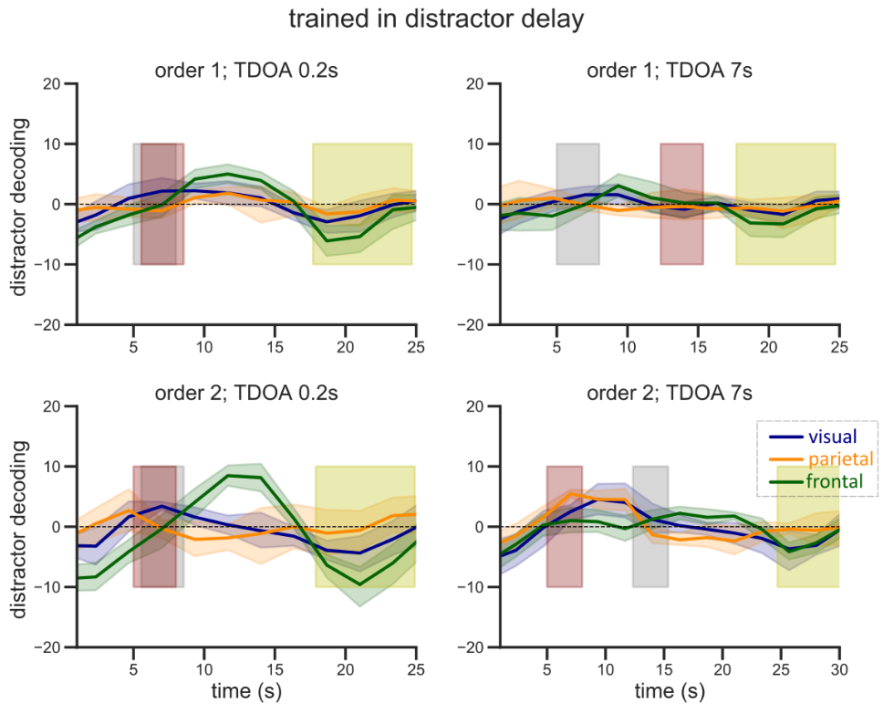

Figure 57 Time evolution of distractor decoding when training in the distractor delay period
Distractor decoding in all four temporal conditions when training in the *distractor delay period*. Boxes represent the different events: grey box represent the presentation of the target; the red box represents the presentation of the distractor and the yellow box the response time. All the events consider a 4s delay due to the hemodynamic response and present a window of 3s to correct for individual variability. The three lines show the mean decoding of the six participants with a 95% ci in the three ROIs (visual, parietal, and frontal areas).

Although distractors are irrelevant information that should be ignored, results clearly show that they can be decoded form different regions, so they are not filtered out at some stage of the processing. While Figure 56 show the time evolution of the decoding of the distractor when training in the *target delay period* (first delay of *order 1 – long TDOA*), Figure 57 shows the same time evolution but for the other training protocol: training in the *distractor delay period* (first delay of *order 2 – long TDOA*).

The results of the previous section showed that the studied visual and parietal areas could not reconstruct the angles of the remembered information when training the IEM in the *distractor delay period* (Figure 55). This result opened two different hypotheses to explain this lack of decoding: first, maybe visual and parietal regions present two different patterns of BOLD signal (one for the target and one for the distractor). Second, maybe visual and parietal decoding rely on sensory strength and training the IEM in relevant information works better than training it in irrelevant information due to attentional-related modulations of sensory cortices (Busse et al., 2008; Hembrook-Short et al., 2017). According to the first hypothesis, as two different patterns of BOLD signal exist (one for the target and one for the distractor), decoding the distractor training in the *distractor delay period* would give a better reconstruction than decoding the distractor training in the *target delay period*. According to the second hypothesis, however, decoding the distractor training in the *distractor delay period* would give worse reconstruction, as the IEM would be better trained in a condition where the sensory cortices had not received an attenuated modulation (*target delay period*). Comparing the time evolution of distractor decoding under the two different training protocols (Figure 56 and Figure 57) and the average decoding during the delay

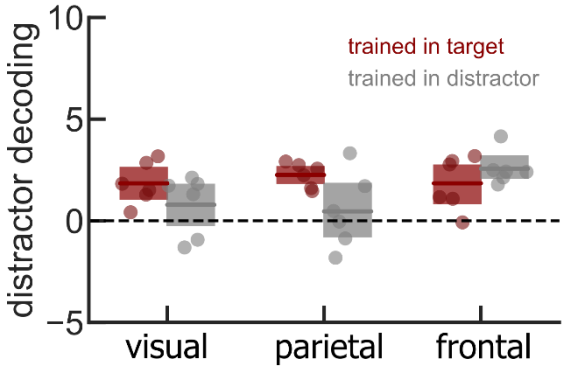

Figure 58 Distractor decoding at different training protocols.
Decoding of the distractor in visual, parietal, and frontal areas when training the IEM either in the *target delay period* and the *distractor delay period.* The box shows the mean decoding of all six subjects and the 95% ci. Each dot is the mean decoding of each subject during the delay period following the distractor presentation.

period following the distractor presentation (Figure 58), visual and parietal regions present better decoding of the distractor when training the model in the *target delay period* than when training in the *distractor delay period* (visual: t=2.29, p=0.02; parietal: t=3.25, p=0.001). No significant difference is observed in frontal (frontal: t=-1.19, p=0.23). Therefore, results support the second hypothesis, suggesting visual and parietal cortices decoding depends on the sensory strength of the stimulus as opposed of holding a different pattern of BOLD signal for the target and the distractor.

Behavioral effects in BOLD signal: attraction-repulsion

Due to the distributed nature of WM (Christophel et al., 2017), I could reconstruct the angles for remembered stimuli in all the regions (Figure 50A). However, these regions may be reflecting some redundant stage of the maintenance process, with no influence on the final response. To address it, I looked for correlations between the actual behavior and the decoding of both the target and the distractor in different regions.

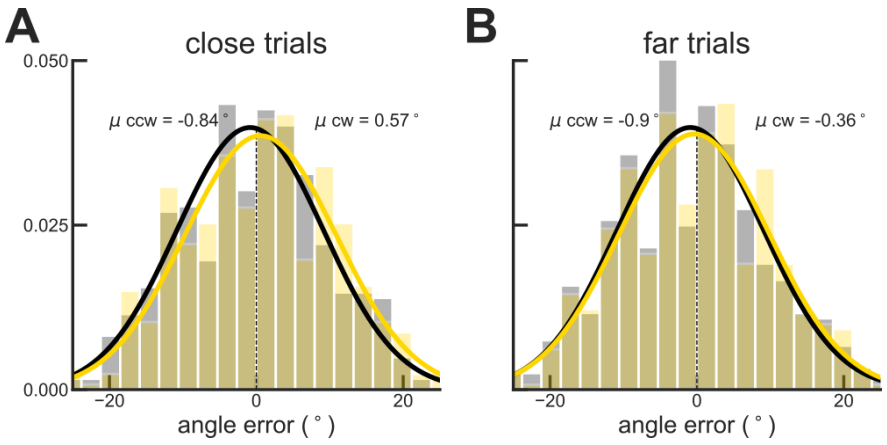

Figure 59 cw-ccw distributions for close and far trial in the scanner dataset

A) Distribution of errors in close trials for clockwise (cw) and counterclockwise (ccw) distractors. A significant difference between the distributions with attractive effect was observed for close trials (mixed linear model, N=6, n=939, β *intercept*= 0.507, z=0.559, ci= [-1.271, 2.285], p=0.576; β *cw-ccw*=-1.444, z=-2.203, ci=[-2.730, -0.159], p=0.028). B) Distribution of errors in far trials cw and ccw distractors. No significant difference between the distributions was observed (mixed model, N=6, n=948, β *intercept*= -0.428, z=-0.515, ci= [-2.059, 1.202], p=0.607; β *cw-ccw*=-0.479, z=-0.733, ci=[-1.759, 0.801], p=0.463).

One of the main behavioral effects I observed was the difference in the interference pattern depending on the distance between the target and the distractor (Figure 47). In *close trials* -distractor located in the same quadrant as the target- I observed an attractive effect (response biased towards the distractor) while in *far trials* -distractor located in a different quadrant - I observed a repulsive effect (response biased away of the distractor). However, when looking at the distribution of errors of cw and ccw trials in the dataset obtained in the scanner (Figure 59), just the attractive effect in the close trials (significant separation between the cw and ccw distributions, p=0.028) was observed.

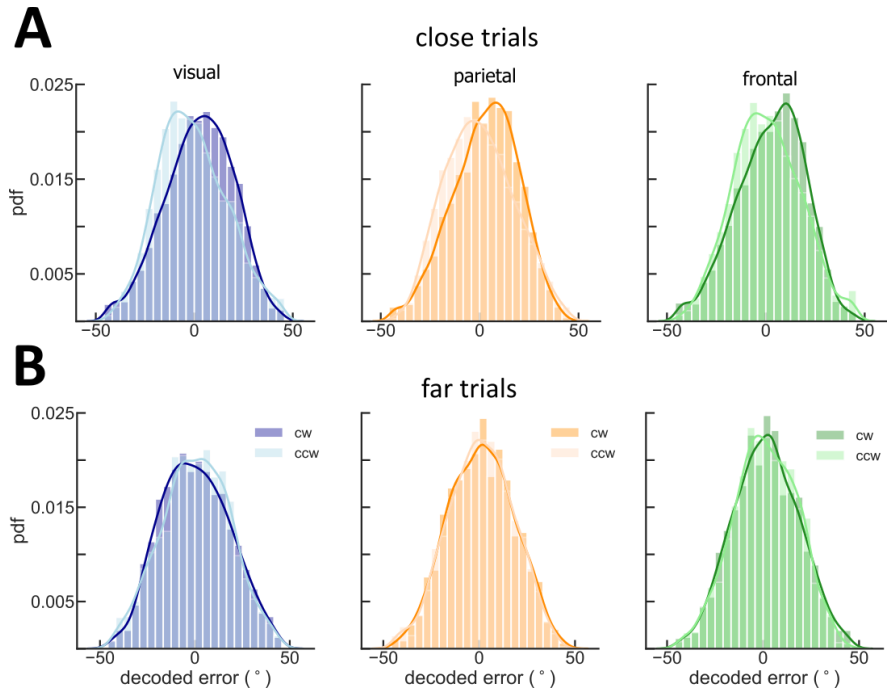

Figure 60 Distribution of reconstructed angles for cw and ccw distractors

A) Distribution of reconstructed angles during the delay period of the close trials. For each distribution, the approximate number of trials was 1887*7 (TRs of the delay period) / 4 (close-far split x cw-ccw split). Significant differences between the cw and the ccw distributions were detected in all three regions (visual: t=8.80, p<0.001; parietal: t=8.03, p<0.001; frontal: t=5.56, p<0.001). B) Distribution of reconstructed angles during the delay period of the far trials. No difference between the cw and the ccw distributions were detected in any of the three regions (visual: t=-0.59, p=0.55; parietal: t=0.53, p=0.59; frontal: t=0.36, p=0.71).

When I used the IEM to reconstruct the remembered angle from BOLD signal from both close and far trials (*Methods-MRI*) and plotted their distributions depending on whether they had a cw or a ccw distractor, a separation between cw and ccw trials was observed in all three regions for close trials (Figure 60A) and no difference was observed for far trials (Figure 60B), consistent with the previously presented distribution of behavioral errors.

Behavioral effects in BOLD signal: TDOA

The lack of differences between regions in the previous analysis may reflect either interference that starts at early stages of the processing that propagates higher up in the hierarchy, top-down feedback, or resolution problems where the distractor was confounded with the target in close trials. For that, I looked for correlations between behavior at the different temporal conditions in the far trials and the decoding of both the target and the distractor in different regions. Figure 48E-F showed the results for both close and far trials of the dataset obtained in the scanner. In both scenarios, I observed a TDOA effect, with higher interference for the short TDOA trials (close trials: β *TDOA*=-0.387, z=-2.453, ci=[-0.696, -0.078], p=0.014; far trials:; β *TDOA*=0.410, z=2.653, ci=[0.107, 0.713], p=0.008). As the decoding procedure tried to decode the isolated stimulus, I focused on far trials (Figure 48F). The mixed model in this condition revealed an interaction order-TDOA (β *orderxTDOA*=-0.238, z=-2.438, ci=[-0.430, -0.047], p=0.015). The subsequent analysis of analyzing TDOA effects in each order condition revealed a repulsive effect for short TDOA in *order 1* and no effect in *order 2* (mixed linear model with random intercept per subject. *Order 1* trials: N=6, n=475, β *intercept*= -0.584, z=-1.053, ci=[-1.672, 0.503], p=0.292; β *TDOA*=0.169, z=2.420, ci=[0.032, 0.307], p=0.016. *Order 2* trials: N=6, n=473, β *intercept*= 0.795, z=0.831, ci=[-1.080, 2.671], p=0.406; β *TDOA*=-0.061, z=-0.915, ci=[-0.191, 0.070], p=0.360). If one of the studied regions is responsible for the maintenance leading to the final response, it must show a decoding pattern compatible with behavioral results.

According to behavioral results, a region responsible for the final memory response should present reduced decoding of the target in those conditions with high interference (*short TDOA*), increased decoding of the

distractor in those conditions with high interference, or both. The exact pattern of target and distractor decoding will also help to describe whether distractor filtering is accomplished through mechanisms that increase the fidelity of the remembered information (Fischer & Whitney, 2012), through mechanisms of distractors suppression (Bettencourt & Xu, 2016) or through a combination of them. All these scenarios predict reduced decoding of the target for higher distracting conditions, either for an active reinforcement of the representation of the remembered angles in the low distracting condition or because of memory interference in the

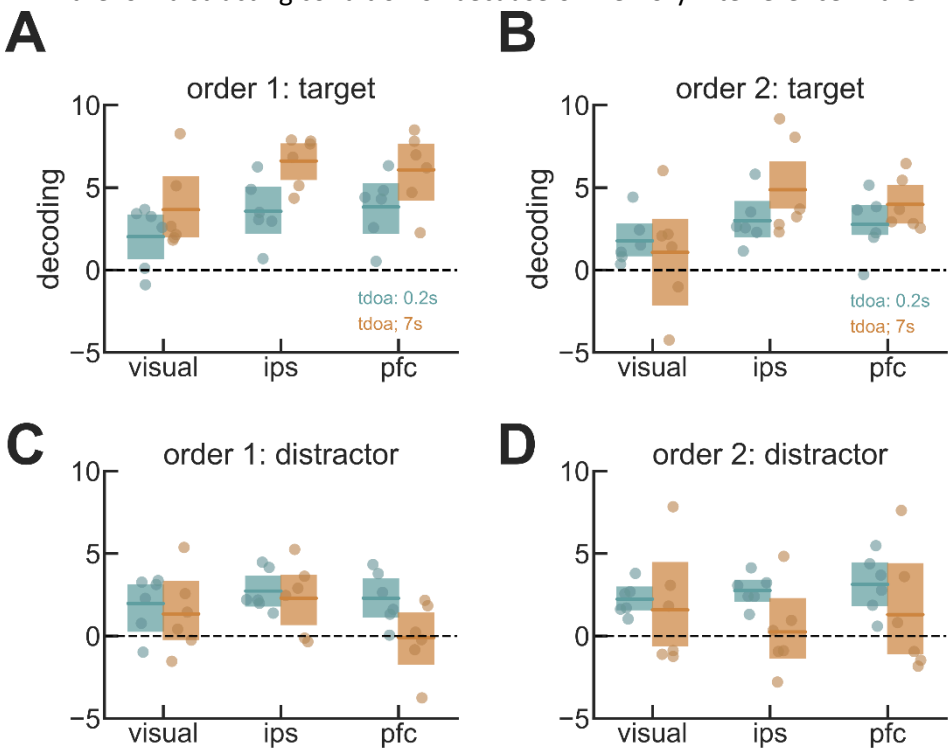

Figure 61 TDOA differences in decoding for the target and the distractor

A) Decoding of the target in visual, parietal, and frontal areas when training the IEM in the *target delay period.* The box shows the mean decoding of all six subjects and the 95% ci. Each dot is the mean decoding of each subject during the delay period following the distractor presentation. Parietal and frontal regions present attenuated decoding of the target in the condition *order1-short TDOA.* **B)** Decoding of the target in visual, parietal, and frontal areas when training the IEM in the *target delay period*. No region presented differences No attenuated decoding at different TDOAs is observed in *order 2* condition. ***C*)** Decoding of the distractor in visual, parietal, and frontal areas when training the IEM in the *target delay period*. Just frontal regions present attenuated decoding of the distractor in the condition *order1-short TDOA.* **D)** Decoding of the distractor in visual, parietal, and frontal areas when training the IEM in the *target delay period*. Just parietal regions present attenuated decoding of the distractor in the condition *order2-long TDOA.*

high distracting conditions. Decoding of the target showed that parietal and frontal regions presented impaired decoding in *order 1-short TDOA* (Figure 61A; visual: t=-1.65, p=0.11; parietal: t=-3.038, p=0.004; fontal: t=-2.10, p=0.042). The decoding of the target was not reduced between TDOAs conditions in *order 2* trials for any region (Figure 61B; visual: t=0.775, p=0.44; parietal: t=-1.94, p=0.06; fontal: t=-1.43, p=0.16). Results regarding the decoding of the target were in line with behavior, and pointed towards parietal or frontal as candidate regions for storing the WM readout. I then evaluated the decoding of the distractor, to evaluate if one of regions presented a more consistent decoding pattern of both the target and the distractor. When decoding the distractor, it is important to clarify that I also used the model trained in the *target delay period*, as it was the one that gave better decoding results of the distractor in visual and parietal regions (Figure 58). Figure 61C-D shows the decoding of the distractor in *order 1* and *order 2* conditions. In *order 1* trials, where I observed a significant effect of TDOA in behavior, just frontal regions showed an increased decoding of the distractor in the short TDOA condition (Figure 61C; visual: t=0.71, p=0.48; parietal: t=0.47, p=0.64; fontal: t=2.25, p=0.029). In order 2 trials, where no significant effect of TDOA was observed in behavior, parietal regions showed an increased decoding of the distractor in the short TDOA condition (Figure 61D; visual: t=0.74, p=0.47; parietal: t=2.58, p=0.014; fontal: t=1.73, p=0.09). Results indicate that frontal area is the one with a decoding pattern, of both the target and the distractor, more consistent with the observed behavior. This suggests the final memory readout is maintained in frontal ares, where the target and the distractor eventually coexist in the same circuit.

Mechanistic explanation for distractor filtering

In this section, I provide a mechanistic explanation of how WM maintenance and distractor filtering are accomplished. The analysis of both behavioral data and BOLD signal provided consistent results with the final memory readout originating in frontal areas, coinciding with the region where the bump attractor model was conceived. To explain the behavioral results, I implemented the bump attractor model combining PA with STP mechanisms (*Methods-Computational modeling,* Network model for distractor filtering), in line with Kilpatrick (2018), Seeholzer et al. (2019) or Barbosa et al. (2020). The model aimed to qualitatively reproduce the observed behavior (Figure 47) under the manipulated distractor features:

Similarity domain:

a) **Distance**: attraction for close distractors and repulsion for distance ones (Figure 46)

Temporal domain:

b) **Order**: increased absolute interference for *order 1* compared to *order 2*.
c) **TDOA**: increased absolute interference for *short TDOA* compared to *long TDOA*.
d) **Order-TDOA**: tendency for an interaction in close trials.

The model qualitatively replicated behavioral results in close trials (Figure 62A). First, the model shows the observed overall attraction for close-by distractors (Figure 47A). The attractive regime is originated due to the ring structure of the model, where neurons coding for close-by locations are more strongly connected than neurons coding for distant ones. This creates an overlap in the activity profile of the memories of the target and the distractor that could lead to the merge of memory traces by the end of the delay. Figure 62B shows a simulated trial of the *order 1 – short TDOA* condition. This condition was the one presenting higher interference, which is reflected in the simulation through the merging of the memory traces, leading to a final position of the bump -the memory readout- located in between the original positions of the target and the distractor.

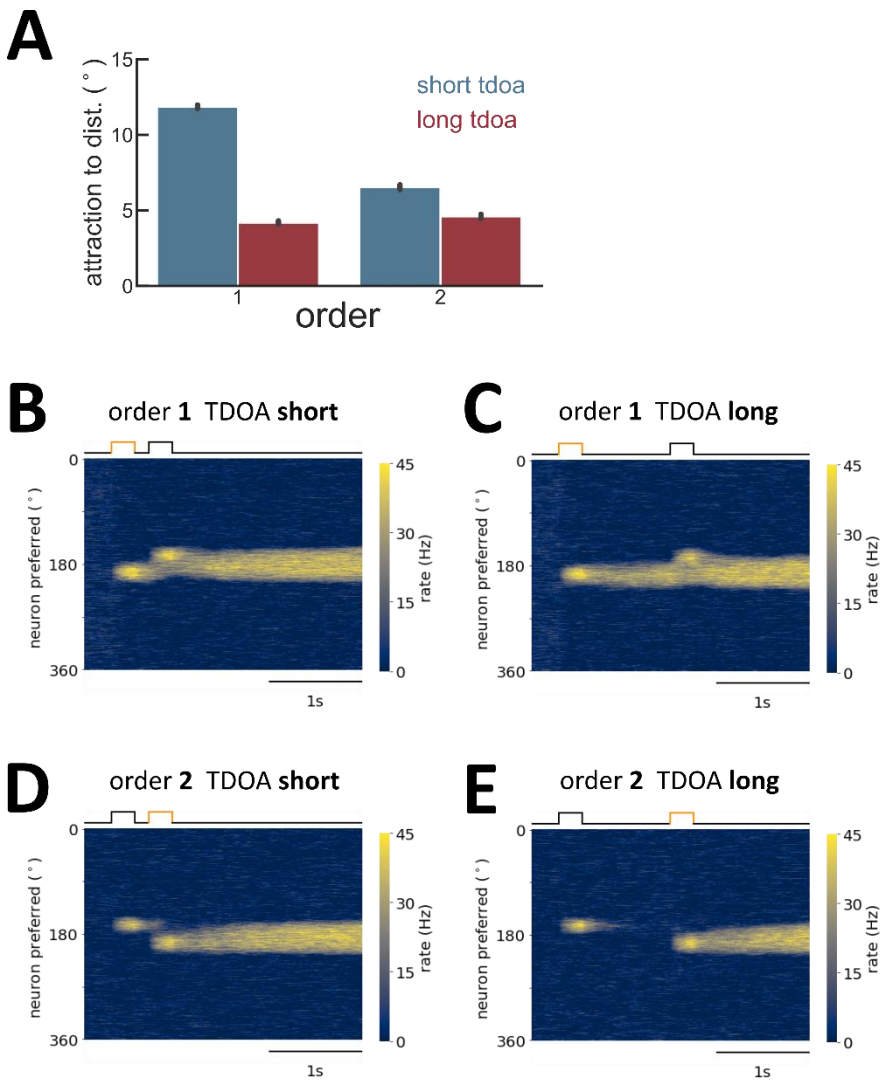

Figure 62 The bump attractor model explains behavior in close trials.

A) Simulated behavior for the different temporal conditions in close trials. An overall attraction was observed in all the condition and an interaction between order and TDOA, having short TDOA a larger interference effect in order 1 compared to order 2. **B)** Simulated trial in the condition *order1-short TDOA*. The simulation shows the rate of each of the 512 excitatory neurons of the circuit and its evolution during a delay of 3s. On top of the simulation there is a line indicating the stimulus presentation (orange) and the distractor presentation (black). The memory readout is computed with a population vector that extracts the mean location of the bump at the end of the delay period. **C)** Simulated trial in the condition *order 1- long TDOA* **D)** Simulated trial in the condition *order 2- short TDOA*. **E)** Simulated trial in the condition *order 2- long TDOA*

The model also replicated the effects of order and TDOA (Figure 47A). First, the *order 1 – short TDOA* was also the condition with higher distraction and, for each order condition, the short TDOA condition had stronger interference effects (Figure 62A). The model showed a clear interaction order-TDOA, which was only observed as a trend in behavioral data (*orderxTDOA*: p=0.067). Figure 62B-E show simulated trials in each of the temporal conditions. The observed effects were obtained thanks to the combination of PA with STP mechanisms. On the one hand, transient STD after the stimulus presentation explained TDOA differences in order 1 (Figure 62B-C): The initial STD generates an unstable state of the bump at the early delay that leads to higher interference for short TDOAs. When the distractor is presented just after the target offset (*short TDOA*), it interferes with a less robust memory trace (attenuated rate) than when the distractor is presented long after the target offset (*long TDOA*), reproducing the observed behavioral pattern. On the other hand, STF explained interference in order 2 conditions (Figure 62D-E): distractors left a synaptic trace that was still present by the time the target appeared. As synaptic traces decay with time, they were stronger by the time the target was presented in the *short TDOA* condition compared to the *long TDOA* condition, leading to more interference in the first scenario. The attractive effect of the synaptic traces, however, is smaller than the one originated by the merging of bumps. Therefore, the magnitude of the interference for short TDOA conditions was larger in order 1, originating the interaction order-TDOA.

Regarding the far trials, the model also replicated behavioral results in far trials (Figure 63A). First, the model shows the observed overall repulsion for distant distractors (Figure 47B). The repulsive regime is again originated due to the ring structure of the model, where the combination of the tuning of the excitatory and the inhibitory connections results in an effective inhibition between memory traces located far from each other. Compared to close trials (Figure 62), the magnitude of the repulsive effects is smaller than the attractive ones, which is also appreciable in behavior (Figure 47). The model replicated just some of the effects of order and TDOA. In behavior, I just observed an effect of TDOA (TDOA: p=0.034), which is successfully reproduced by the model. However, it also predicted significant effects of order and of the interaction order-TDOA. More simulations for the far conditions are still needed to understand the differences between the behavior and the model. Differences could be

originated by the sampling of close and far trials in behavioral data: while close trials explored a range of 20° (10°-30°), the range for the far trials was 140° (40°-180°).

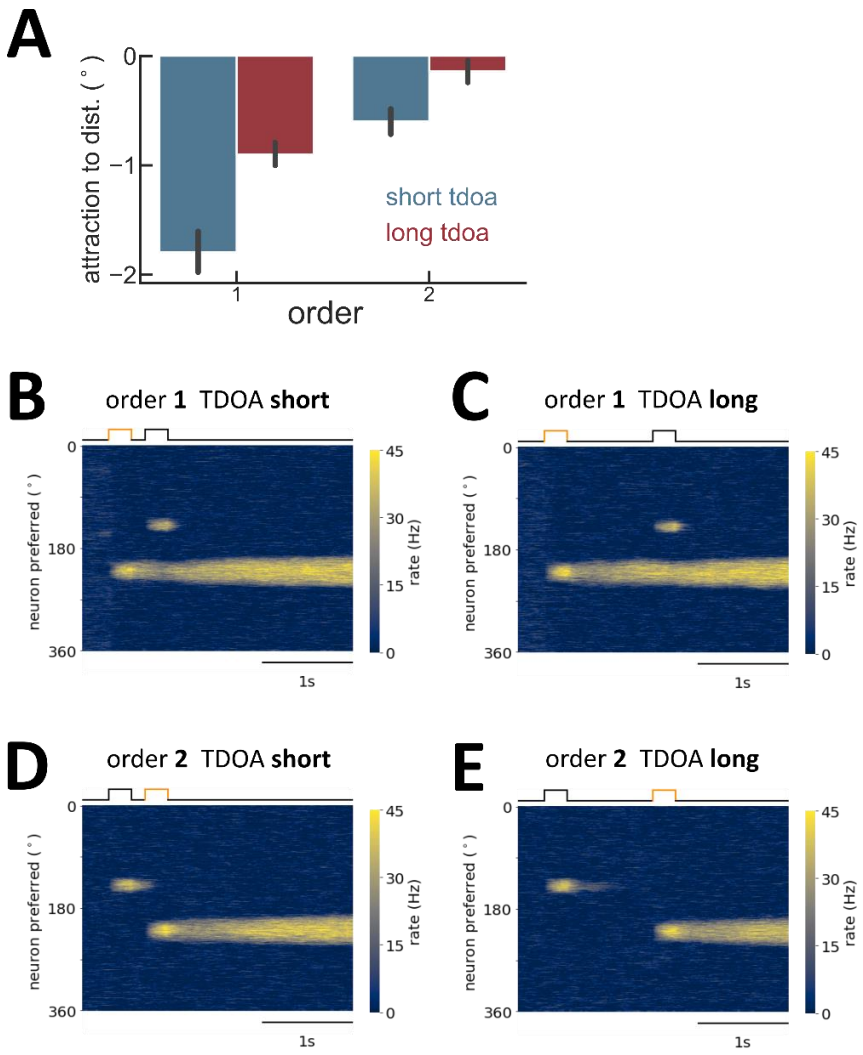

Figure 63 The bump attractor model explains behavior in far trials.
A) Simulated behavior for the different temporal conditions in far trials. An overall repulsion was observed in all the condition and an interaction between order and TDOA, having short TDOA a larger interference effect in order 1 compared to order 2. **B)** Simulated trial in the condition *order1-short TDOA*. **C)** Simulated trial in the condition *order 1- long TDOA* **D)** Simulated trial in the condition *order 2- short TDOA*. **E)** Simulated trial in the condition *order 2- long TDOA*

Distractor filtering: electrophysiology

Through a collaboration with Jaqueline Gottlieb, I got access to neural recordings of monkeys performing a vsWM task with prospective distractors at different TDOA while being recorded in LIP and dlPFC (*Methods-Electrophysiology*, Dataset 2). This dataset was used for Suzuki & Gottlieb (2013), where they observed that distractor responses were more strongly suppressed and more closely correlated with performance in the dlPFC relative to LIP. Besides pointing to frontal regions as a candidate area for the final memory readout, this dataset allowed me to address two different questions. First, if the population code underlying the representation of the target was stable or dynamic over time, and second, if the rate activity profile of dlPFC neurons was the one predicted by the computational model.

Previous studies using cross-temporal decoding found differences in the stability of the code for the stimulus presentation and the delay period (Mendoza-Halliday & Martinez-Trujillo, 2017; Spaak et al., 2017; Stokes et al., 2013). When I trained and tested a linear decoder across all the time intervals in the different TDOA conditions (Figure 64), I observed a similar pattern of cross-temporal generalization: stable code throughout the stimulus presentation and stable code throughout the delay period (Figure 64A-D). The significance of each train and test bin (Figure 64E-H) was evaluated with a permutation test comparing the mean decoded error with a distribution of shuffled reconstructions -shuffling the labels of the testing subset-. Comparing the stability of the code during the delay period for the different TDOA condition, larger errors were detected for short TDOA conditions compared to long TDOA conditions (Figure 64A-D). A subsequent analysis of the decoding strength at this interval (mean error per neuron for training-testing times>800ms) showed better reconstructions for longer TDOA conditions, in line with behavioral results (Figure 65, linear mixed model with a random intercept per neuron: n=63, β *intercept*=58.807, z=35.819, ci=[55.589, 62.025], p<0.001, β *TDOA*=-0.006, z=-6.773, ci=[-0.008, -0.004], p<0.001).

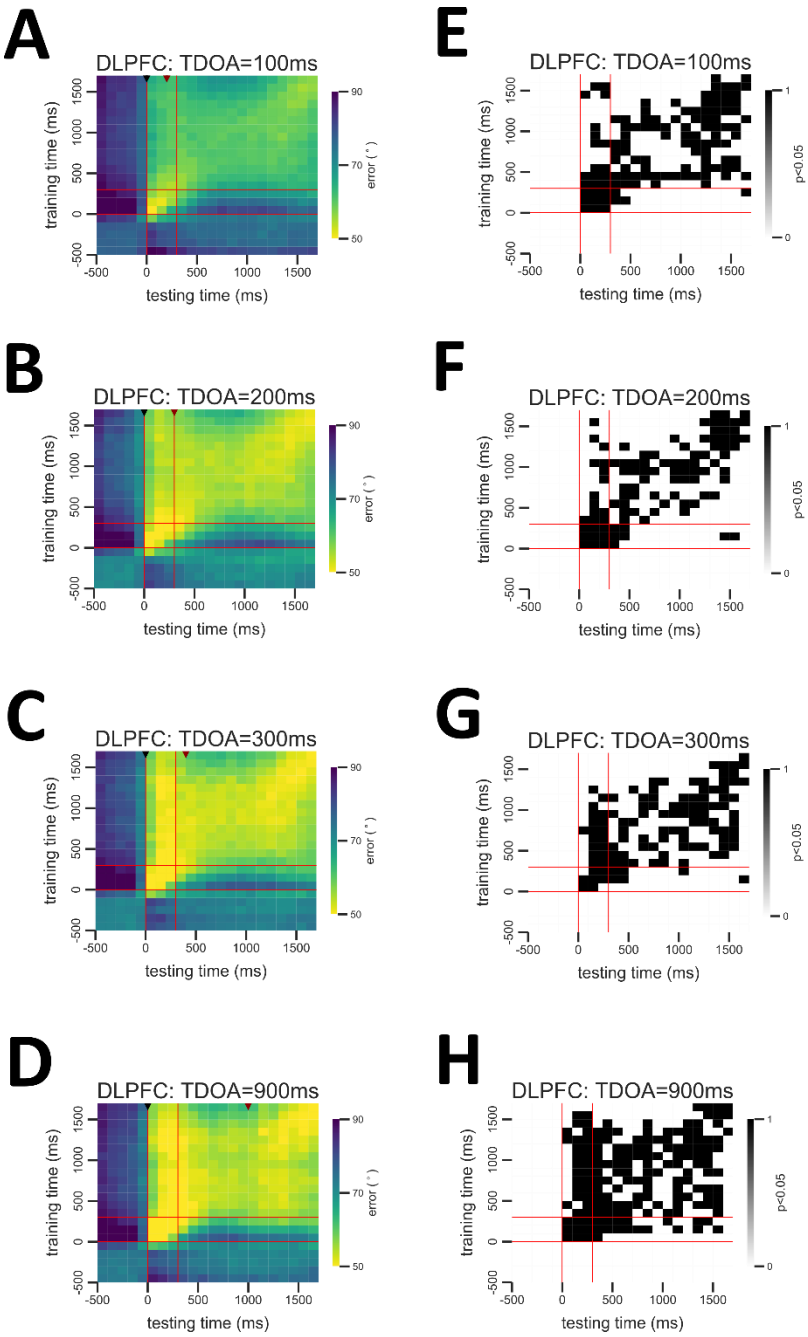

Figure 64 Code stability for different TDOAs

A-D) Cross-temporal decoder for the different TDOA conditions. The y axis shows the training time and x axis the testing time. The color bar illustrates the error of the decoding, so brighter colors indicate better reconstruction of the target. The black small triangle indicates the presentation time of the target, and the red one, the presentation time of the distractor. **E-H)** Significance of each cross-decoding bin evaluated with a permutation test (mean error of all neurons compared to the distribution of shuffled errors). Black squares show p-values <0.05.

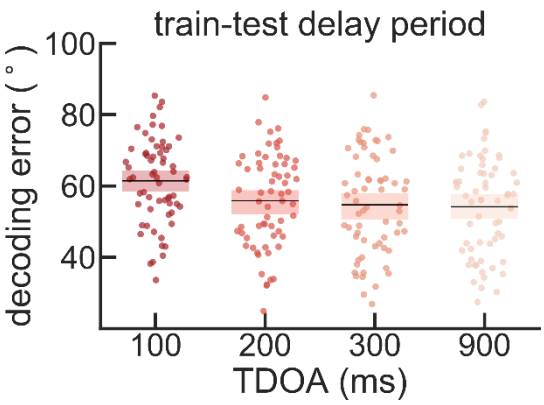

Figure 65 Delay cross-temporal error reduction with TDOA
Mean error for training and testing times larger than 800ms for the different TDOA conditions. Each point represents the mean error of each neuron (n=63), and the box shows the population mean and the 95% ci. A significant reduction of the error is observed for long TDOA conditions (mixed model, *TDOA* p <0.001).

When computing the mean cross-temporal reconstruction for the whole region (Figure 66A, collapsing TDOA conditions), I observed an important difference with previous results regarding the symmetry of the cross-temporal generalizations: while training during the stimulus presentation compromised decoding during the delay period (same as previous studies), training at the end of the delay period allowed decoding during the stimulus presentation (Figure 66B).

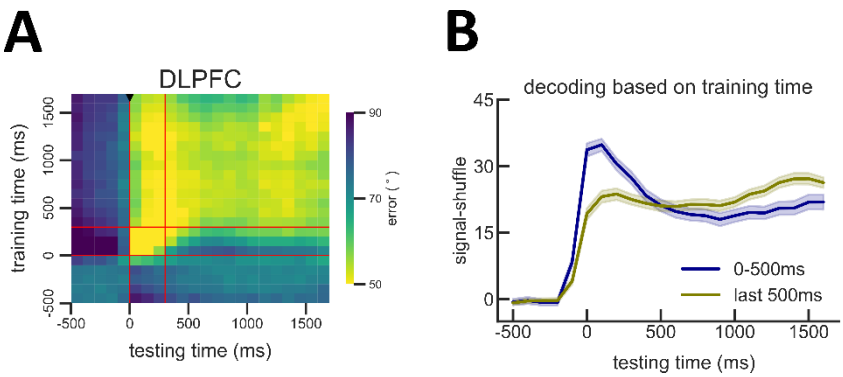

Figure 66 Average cross-temporal decoding
A) Mean cross-temporal decoding of the target in PFC. **B)** Testing accuracy when training during the stimulus presentation (blue) or during the last 500ms of the delay period (olive). The y axis shows the difference between the signal and the shuffled cross temporal decoding matrixes for the selected training and testing times (computed individually at each TDOA condition) and flipped (multiplied per -1). The x axis shows the testing times. While training during the stimulus presentations compromises decoding during the delay period, training at the end of the delay period does not compromise decoding during stimulus presentation.

The computational model presented in the previous section made a clear prediction regarding the activity profile of the neurons responsible for the final memory readout: stimulus selectivity followed by a dip of activity due to STP and a following recovery of the rate (Figure 14). This dataset also allowed me to test the firing profile predicted by the computational model by correlating the decoding strength of each neuron with the presence of a transient dip after the stimulus offset. To do so, I analyzed the firing activity the firing activity of the recording neurons when the target was presented in the RF (Figure 67A) and fitted a parabolic curve ($y = ax^2 + bx + c$) to each of them. Figure 67B shows examples of two neurons: the one on the left presenting a parabolic firing profile (a=0.52) and the one on the right a linear profile (a=-0.05). The computational model predicted that neurons with a parabolic profile are the ones responsible for the final memory readout, so I correlated the decoding error from the cross temporal decoder with the a of the parabolic fit (Figure 67C) and indeed observed a significant correlation (linear model, n=252, β *intercept*=57.1, z=39.17, ci=[54.226, 59.967], p<0.001, β a=-6.47, z=-2.821, ci=[-10.987, -1.953], p=0.005), where neurons with a pronounced dip in the firing rate at early stages of the delay presented a better final decoding of the target.

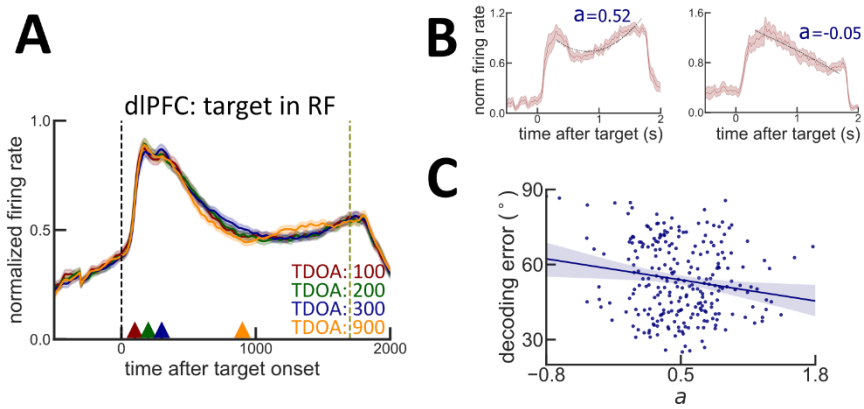

Figure 67 Correlation between electrophysiological rate profile and decoding strength

A) Normalized firing rate of neurons from Suzuki & Gottlieb, (2013) where the target was presented in the RF of the neurons and the distractor was not. Each line shows the mean firing rate of 63 neurons at the different TDOA conditions. The dashed black line shows the time of the target presentation and the yellow one the response time. Each triangle shows the presentation time of the distractor. **B)** Two examples of the parabolic fit on top of the firing profile of a single neuron. The a parameter of the fit is displayed, with larger values of a for pronounced dips during the delay period. **C)** Correlation between the cross-temporal decoding strength of each neuron during the last 500ms of the delay with the estimated a of the parabolic fit.

Distractor filtering under NB stimulation

In this section, I wondered if the bump attractor computational model could provide a mechanistic explanation of distractor filtering for the electrophysiological and behavioral results of Qi et al. (2021). In that work, two implanted monkeys performed a vsWM task with distractors while stimulating the Nucleus Basalis of Meynert (NB). Similar to the paradigm used in the section *Distractor filtering: TDOA and order effects* (*Methods-Paradigms and analysis*), distractors were manipulated in the similarity (target-distractor angular distance) and temporal domain (*Methods-Electrophysiology*, Dataset 3). However, in the temporal domain, just *order* was manipulated. To avoid confusions when consulting the original source, I have maintained the nomenclature of the Qi et al. (2021), which is *Remember 1st* (equivalent to *order 1*) and *Remember 2nd* (equivalent *order 2*). Instead of using an interleaved-trial design, the experimenters wanted to minimize the uncertainty about the trial type, so they used a block design with 10 consecutive trials of *Remember 1st* followed by 10 consecutive trials of *Remember 2nd*. Thus, the bump attractor model had to explain the observed behavioral and electrophysiological results with and without stimulation of the NB with a control strategy consistent with the block design.

Considering the task design, I implemented a bump attractor model that could maintain the WM content in two different regimes: *Remember 1st* and *Remember 2nd* through a neuromodulation in the conductances (*Methods-Computational modeling*). Consistent with known cholinergic activation of PFC neurons (Carr & Surmeier, 2007; Hedrick & Waters, 2015), I modeled NB stimulation as an unspecific increase of excitability of excitatory neurons in the circuit through a slight increase in external input. This provided two independent axes to test the performance of the network model: task condition (*Remember 1st*/*2nd*) and NB stimulation (ON/OFF). To evaluate the performance of the network, I compared the readout of the network at the end of the delay with the location of the target to obtain a measure of behavioral error. I then inspected if the computational model could reproduce the results of Qi et al. (2021) and thus provide evidence towards attractor dynamics in PFC controlling distractor interference in WM.

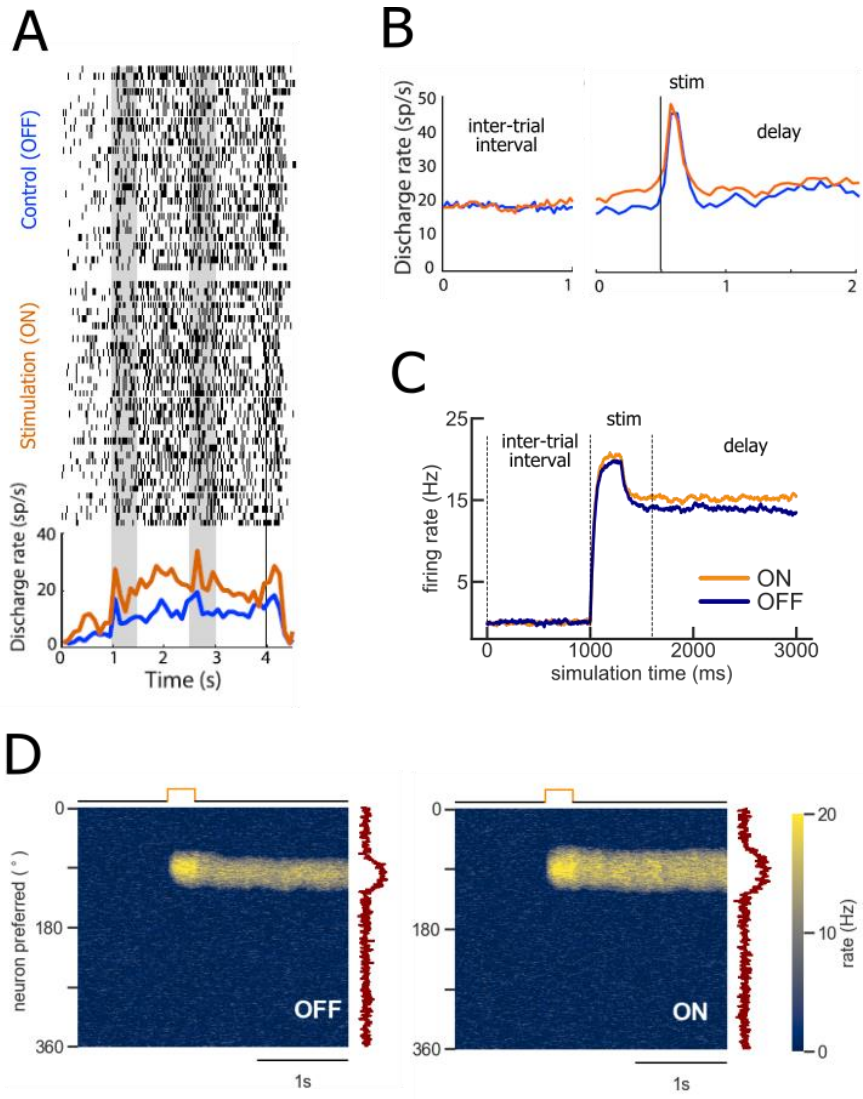

Figure 68 The bump attractor model reproduces the increase of firing rate during the delay period under NB stimulation in PFC.

A) (Modified from Qi et al. (2021)) Raster plot of a PFC neuron with and without NB stimulation (ON and OFF). **B)** (Modified from Qi et al. (2021)) Mean activity of the 54 neurons that responded to visual stimulus and presented elevated firing during the fixation period after NB stimulation. Enhanced firing rate in the ON condition is observed through the delay period but not in the ITI. **C)** Average firing of 100 model neurons responding to preferred stimulus in bump attractor model simulations in the OFF and ON condition. **D**) Two network simulation examples of the bump attractor model in the OFF (left) and ON (right) conditions. NB stimulation elicited elevated firing rate during the delay period as well as a broadening of the bump (compare bumps in average network activity at the end of the delay, red traces).

Qi et al. (2021) analyzed the activity of 54 stimulus-responsive neurons that reacted to NB stimulation with elevated firing rate during the fixation period. In those neurons, they observed that NB stimulation induced elevated firing during the delay period (Figure 68A-B) but not during stimulus presentation or in the inter-trial interval (ITI). The computational model displayed qualitatively similar neural dynamics. Figure 68C shows the average peri-stimulus firing rate in model neurons responding to preferred stimuli in the ON and OFF conditions. The model qualitatively reproduced electrophysiological results, with no appreciable difference between the OFF and ON conditions during the ITI, small differences during the stimulus presentation and clear differences in the firing rate during the delay period. Full network simulation examples in the OFF and ON NB stimulation conditions can be observed in Figure 68D.

A broadening of the bump is noticeable (Figure 68D, in red), consistent with electrophysiological results of Qi et al. (2021), and which will be important for subsequent points. In sum, the activity of individual neurons in the bump attractor network is modulated by a slight increase in cellular excitability (simulating NB stimulation) in a qualitatively similar way to the effects of NB stimulation on the activity of PFC neurons (Qi et al., 2021).

I then analyzed whether this network could account for the behavioral effects observed experimentally by Qi et al. (2021) (Figure 69). Their paradigm consisted in two different blocks: *Remember 1st* and *Remember 2nd*, depending on which of two sequentially presented stimuli was to be remembered and reported at the end of the trial to obtain reward. Also, the distance between these two stimuli was manipulated (*Methods-Electrophysiology*). I tested the computational model specifically in these various conditions: *Remember 1st*/*2nd*, distractor close/far, and NB stimulation ON/OFF are examples of simulated trials in all these conditions. In the *Remember 1st* condition (Figure 69, left), the computational model reproduced the behavioral effects observed in Qi et al. (2021) (Figure 69C: modeling results; Figure 69E: original behavioral results). Performance was impaired in the ON condition (NB stimulation), specifically when the distance between the target and the distractor was small (close trials), but performance improved when the distance between the target and the distractor was large (far trials). Increased firing rates in the ON condition makes more neurons to be engaged in the bump activity, which results in a broader bump. This broader bump -compared to the OFF

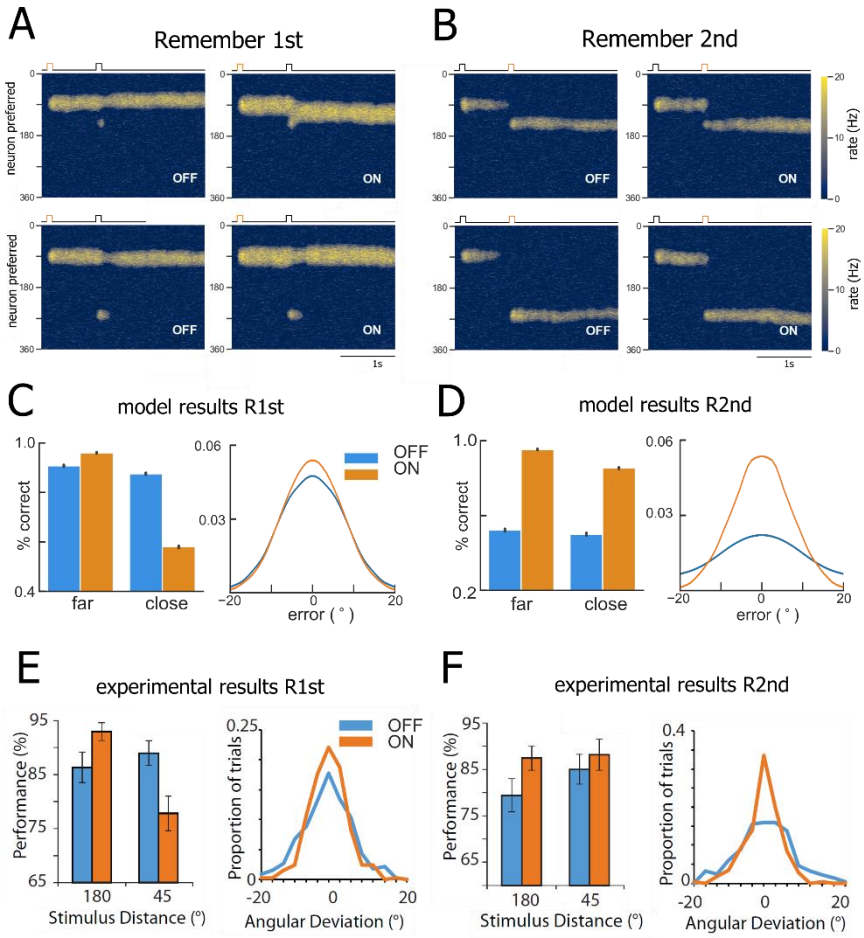

Figure 69 The bump attractor model reproduces the observed experimental performance.

A) Sample simulation showing activity of excitatory neurons in a bump attractor network for both the OFF and ON conditions in *Remember 1st*. Abscissa represents time and ordinate represents neurons with preference for different visual stimulus locations, indicated by location varying between 0 and 360°. Activity of neurons with different preferences is indicated based on color scale. The first visual stimulus appearance at 90° elicits a bump of activity that is maintained during the delay period, after the visual stimulus is no longer present. In the top left panel, I simulated the appearance of the second visual stimulus at the 150° location (OFF, close). As the simulated response is the readout of the final location of the bump at the end of the delay period, the appearance of the second stimulus does not disrupt the initial bump, as the final position of the bump is nearly the same as the initial one (small variations due to noise fluctuations). On the other hand, The ON condition (top right) results in a broader bump of activity. Therefore, when the second visual stimulus appears in a nearby location, it is not inhibited by lateral inhibition and the bumps are more likely to merge, compromising behavior. The second row represents the far conditions, where the second stimulus is located at 260º, away from any possible interference with the initial stimulus. In this scenario, the ON condition again results in a broader bump of activity, which is more resistant to noise fluctuations, thus improving behavior.

B) The modulated network for the *Remember 2nd* regime also reproduced behavioural results. In the *Remember 2nd* condition, the initial bump of activity terminates, and a new bump is maintained after the appearance of the second visual stimulus. For the remember second condition, the same improvement observed for the far condition in remember first occurs with NB stimulation: broader bumps are more resistant to noise fluctuations, so they diffuse less during the delay period leading to more accurate responses. In this case, as the bump from the first stimulus is already extinguished, there is no effect of bump interference and there is an overall improvement of performance independent of the distance between stimuli. **C & D**) Show the model's simulated behaviour. **C**) In the *Remember 1st* (R1st) condition, the model showed improved performance for the far condition and impaired performance for the close condition for the ON condition (left). Results of 20,000 simulated trials are shown for each condition with a 15º threshold for correct trials. The right plot shows the distribution of simulated errors in the far condition. Simulated responses are more accurate in the ON condition due to the enhanced noise resistance of the broader bumps. **D)** In the *Remember 2nd* condition (R2nd), the model showed an overall increase of performance for the ON condition (left). The distribution of the simulated errors showed more accuracy for the ON condition due to enhanced noise resistance and extinction resistance. **E & F**) Show the actual experimental data for the *Remember 1st* and *Remember 2nd* conditions. In both scenarios, the experimental data is qualitatively reproduced by the model (directly on top). **E**) (Modified from Qi et al. (2021)) Mean monkey performance (and sem) in the *Remember 1st* condition (left), for trials grouped by distance between the first and second visual stimulus. The right plot shows the empirical distribution of angular deviations from mean saccadic endpoint for the far condition. **F)** (Modified from Qi et al. (2021)) Mean monkey performance (and sem) in the *Remember 2nd* condition, for trials grouped by distance (left). The right plot shows the empirical distribution of angular deviations from mean saccadic endpoint for the far condition.

condition- is more easily attracted to a distractor activation located at a close distance on the network, resulting in an attraction effect that compromises performance (Figure 69A, top). On the other hand, for those trials where the distractor was located far away, the increased broadening of the bump had positive effects for WM performance. A broader bump leads to a more stable bump attractor, less sensitive to noise fluctuations in the network, thus leading to reduced bump diffusion during the delay (Wimmer et al., 2014) and more accurate read-outs at the end of the delay (Figure 69C, right).

When no distractor is located close enough to interfere with the bump reduced bump diffusion is the primary effect of stimulation, and performance improves (Figure 69A, bottom). In sum, a broadening of bump attractors caused by NB stimulation can explain WM impairment in close-distractor trials and WM improvement in far-distractor trials in the *Remember 1st* condition, as observed experimentally (Figure 69E, Qi et al. (2021)).

The model also reproduced the effects in the *Remember 2nd* regime (Figure 69D, right). In this condition, Qi et al. (2021) found an overall performance improvement after NB stimulation (Figure 69F). Again, the bump attractor computational model could explain this behavior with the same mechanistic explanation of broader bumps after NB stimulation. In this case, as the to-be-remembered stimulus is the second one, the interference with the distractor in the close-distractor trials is strongly reduced, so the effect that prevails in all cases is the reduced random diffusion of the bump during the delay period due to bump broadening. In sum, slight depolarization in the ON condition (simulating NB stimulation) caused broader bumps and reduced bump diffusion in Remember-2nd network simulations, which can explain WM improvement observed for this condition experimentally.

The experimental design contained null conditions where one of the two stimuli was not presented (20% of the trials). Monkeys doing the task could not anticipate these trials, so they were expecting to observe a stimulus that did not appear. In the null conditions *Remember 1st-Absent 2nd* (Figure 70A) no difference was observed during the delay period upon NB stimulation. However, in the condition *Remember 2nd-Absent 1st*, NB stimulation revealed neural responses precisely at the time of the expected, but absent, 1st stimulus (Figure 70B). In the context of bump attractor models, this could reflect the appearance of “phantom bumps” - a bump appearing spontaneously at a random location because of temporal anticipation in this null condition. I used the computational model to test this intuition. I modeled the presentation of the stimulus as the combination of two currents: a non-specific input temporally aligned to the stimulus onset representing a timing anticipatory signal (i.e. temporally specific but stimulus non-specific) and a temporally and stimulus specific input representing the visual stimulus. Simulated trials reproduced the electrophysiological results: for *the Remember 1st-Absent 2nd* condition, no phantom bump was observed during the second delay period, as the already present bump inhibited the rest of the circuit (Figure 70C). However, in the *Remember 2nd-Absent 1st* condition, phantom bumps appeared in the ON but not in the OFF condition (Figure 70D). In the OFF condition, the anticipation input was not strong enough to elicit a bump (Figure 70D, left). However, in the ON condition, this anticipation current together with the one resulting from NB stimulation was sufficient to destabilize the un-patterned baseline network activity and elicit

phantom bumps in some trials (Figure 70D, right). Because phantom bumps appear unpredictably at different points in the network, it would be difficult to validate this hypothesis experimentally from the activity of just a few simultaneously recorded neurons.

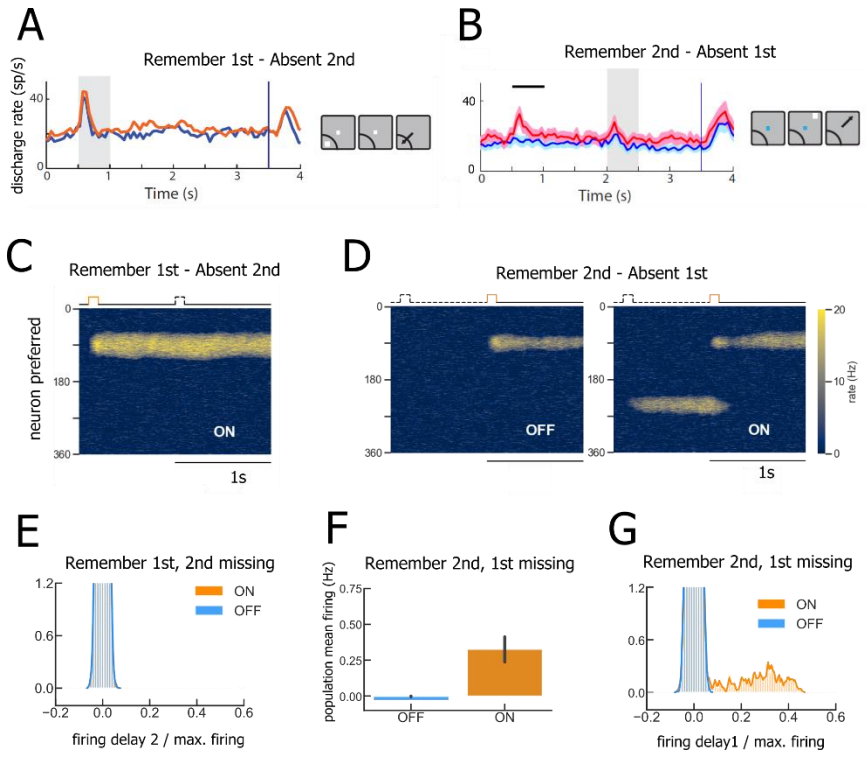

Figure 70 Phantom bumps explained by the model.

A) Mean firing rate in the "null condition" *Remember 1st-Absent 2nd* (modified from supplementary information of Qi et al. (2021), supplementary figure). No specific "response" is observed in the ON condition at the time in which the 2nd stimulus was expected (2s). **B)** Mean firing rate of the two monkeys in the "null condition" *Remember 2nd-Absent 1st* (modified from Qi et al. (2021)). Horizontal lines illustrate the times that a first visual stimulus would have been delivered. Elevated activity during this period can be observed for the NB-stimulation ON condition. **C)** Simulation of a *Remember 1st-Absent 2nd* trial in the ON condition. **D)** Simulation of a *Remember 2nd-Absent 1st* trial in the OFF and ON conditions. **E)** Distribution of simulated rates outside the preferred location during the first delay divided by maximum firing rate of the simulation (preferred location during second delay). Complete overlap for the absent second stimulus condition in both ON and OFF conditions, indicating the absence of "phantom bumps". **F)** Average simulated firing rate following the absent first stimulus, in the ON and OFF conditions. To corroborate this effect was a consequence of the appearance of sporadic bumps and not an overall baseline elevation, I **G)** quantified the mean firing rate in the period following the absent stimulus and divided it against the mean response to a preferred stimulus (phantom/ bump activity). Phantom bumps are revealed by a slight elevation of the histogram for the ON condition. The bump emerged at random locations; thus, any given neuron was activated in only a small fraction of trials.

The fact that the population mean firing rate during the first delay (Figure 70F) was higher for the ON condition, could still be explained based on a small general elevation of baseline activity, without the formation of phantom bumps. I thus devised an analysis that could be tested in single-neuron data. In each trial I normalized the rate of each neuron in the delay following the absent stimulus by its mean firing rate in response to a preferred stimulus (Figure 70E,G). In the *Remember 1st-Absent 2nd* null condition, no difference was observed between the histograms of these normalized rates in the ON and OFF conditions (Figure 70E). However, in *Remember 2nd-Absent 1st* simulations I observed that in the ON condition, some neurons presented elevated activity in a few trials that approached their preferred rates during bump states (Figure 70G), indicating their participation in bump activity. In sum, the bump attractor model explained the emergence of electrophysiological markers of stimulus anticipation under NB stimulation (Qi et al., 2021) as the triggering of phantom bumps by arousal mechanisms in cholinergically depolarized prefrontal networks.

Network simulations allowed me to address an additional question, regarding the site of action of NB stimulation. In principle, the neural effects observed by Qi et al. (2021) could have been entirely the result of changes in upstream, sensory cortical areas that were not active during the delay period of the task. Such upstream changes would then be propagated and maintained in the PFC. I tested this hypothesis in the model by simulating NB stimulation as an increase of the external current just during stimulus presentation (ON exp), and not during the whole trial (ON). Like the ON condition, ON exp also resulted in a broadening of the bump during the stimulus presentation (Figure 71A-B). During the delay period, however, ON exp did not maintain the broader bump (Figure 71C-D), because the bump attractor model imposes a fixed bump width in the absence of selective input during the delay period. In the absence of a broader bump during the delay period, behavioral effects due to interference or diffusion stability cannot be reproduced, suggesting NB stimulation also acts in the memory circuit of PFC besides other sensory areas.

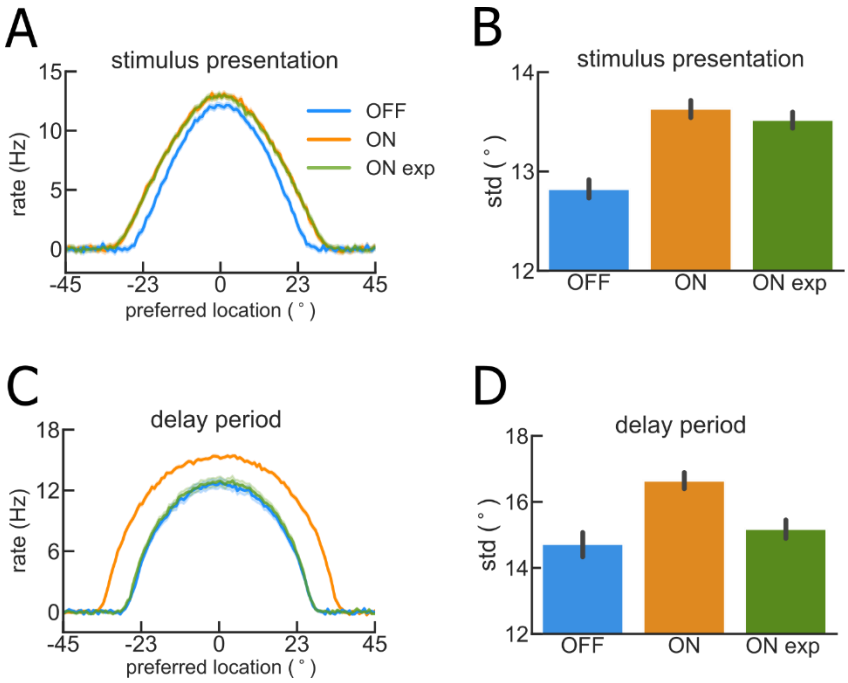

Figure 71 NB stimulation: memory vs sensory-related effect.
Model comparison of NB stimulation as elevated current just during the stimulus presentation (ON exp) vs elevated current during both stimulus presentation and delay period (ON). **A)** Presents the average firing during one second of the stimulus presentation of 100 simulations centered to the preferred stimulus. A broader bump is observed for both ON an ON exp compared to the control condition (OFF). **B)** Mean standard deviation of a gaussian fit of 100 simulations illustrate this difference in bump-broadening. **C)** Same as A but, instead of sampling one second of stimulus presentation, I analyzed the last second of the delay period. In this scenario, the ON exp. the condition did not present a broader bump than the control condition. **D)** Same as B but for the last second of the delay period. To achieve broader bumps during the delay period is not enough with elevated current activity during stimulus presentation (ON exp), this elevated current must also be present during the delay period (ON).

Behavioral and electrophysiological results of Qi et al. (2021) contain a paradoxical effect: while performance is enhanced in many conditions after NB stimulation (Figure 69), neurons in PFC experienced a loss of tuning (Figure 72A). Previous works reported a link between tuning curve properties and behavior (Busse et al., 2008; Compte & Wang, 2006; Li et al., 2004; Raiguel et al., 2006; Sanayei et al., 2018; T. Yang & Maunsell, 2004), suggesting that better tuning (in terms of width or amplitude) leads to better performance. The results presented here, however, show that decrease of tuning at the neuronal level does not always imply an impairment in performance, which is the result of a readout from the

whole network. In the computational model, I computed the final readout as the population vector of all the neurons at the end of the delay period. In the experiments, NB stimulation weakens neuronal tuning through a general increase of firing rates in the tuning curves (Figure 72A). This is very different from the situation in the rate computational model (Figure

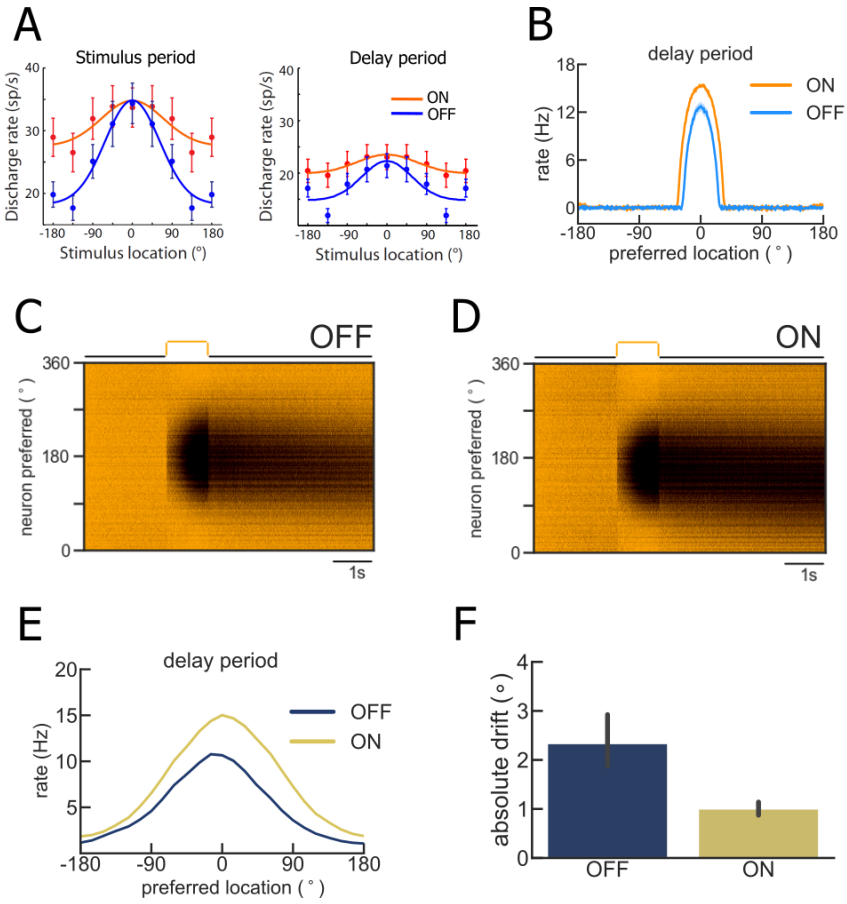

Figure 72 Spiking bump attractor model.

A) Population tuning curves for both the stimulus presentation and the delay period (from Qi et al, (2021)). They are obtained by averaging responses of individual neurons to visual stimuli relative to each neuron's preferred reference location (centered at 0). **B)** Population tuning curve of the rate version of the bump attractor model. **C)** Control simulation in the spiking version of the bump attractor model. **D)** NB stimulation simulation in the spiking version of the bump attractor model. **E)** Population tuning curve of the spiking version of the bump attractor model. **F)** The spiking version of the bump attractor model also presented Increased performance in the NB stimulation simulated condition (ON). The graph shows how diffusion is reduced in the ON condition, leading to more accurate reports at the end of the delay period (2 (OFF/ON) x 25 (stimulus positions) x 5 (repetitions)=250 simulations).

72B), where only responses to preferred stimuli get enhanced in the ON condition, but this is because this is a very simplified model, composed of identical neurons with fixed tuning and connected all-to-all. This scenario is an idealized approximation, so I also tested a more realistic spiking network with strong heterogeneity in neuronal tuning caused by sparse connectivity (Hansel & Mato, 2013).

When I simulated NB stimulation the same way as in the rate model (baseline condition in Figure 72C compared to increase of the external current in Figure 72D) I indeed observed that stimulation increased neuronal responses across the full extent of the tuning curves (Figure 72E), similar to the experimental data (Figure 72A). Finally, I analyzed whether the loss of tuning in the spiking network was associated to impaired or improved performance. The spiking network also presented reduced bump diffusion (increased performance) in the NB stimulation condition (Figure 72F). Overall, modeling results of a more realistic version of the bump attractor model support the interpretation that NB stimulation induces better performance at the population level even when tuning is impaired at the single-unit level.

5.Discussion

Topography of the working memory circuit

In the first section of this chapter (*Angular dimension*), I made a topographical exploration of the vsWM circuit by analyzing biases and precision at different eccentricities. While studying biases and precision in the memory reports is a standard approach when testing capacity (Almeida et al., 2015; Chieffi & Allport, 1997; Nassar et al., 2018; Pratte et al., 2017; Sprague et al., 2014), distractor effects (Chunharas et al., 2019; Rademaker et al., 2015; Van Ede et al., 2018) or serial biases (Barbosa et al., 2020; Stein et al., 2020); their combination with topographical aspects is seldom (W. J. Harrison & Bays, 2018; Staugaard et al., 2016). Therefore, in this section, I studied how interference changes with eccentricity, and I proposed a computational model that mechanistically explains the observed behavior and challenged predictions of the SRT.

I developed a vsWM task with single-item trials and multi-item trials presented at different eccentricities and variable delay lengths. In single-item trials, I found an unexpected increase of accuracy in the angular dimension with eccentricity (Figure 36B). The bump attractor model explained it through a loss of tuning with eccentricity. More eccentric bumps were then broader, absorbing more efficiently the noise fluctuations that cause the delay-diffusion. For the multiple-items trials, I observed an attractive regime for further eccentricities and a repulsive regime for close ones (Figure 37B). Again, the loss of tuning with eccentricity explained these effects, as broader bumps are more prone to attract themselves at fixed azimuthal angles. A limitation of this experiment is that I tested three fixed angular distances, so a larger exploration of the spectrum would help to validate the model in a parametric way. In a similar experiment, Almeida et al., (2015) controlled for the distances between items and observed a reduction in interference effects, providing a nice validation of the spatial properties of the model. Future experiments testing different eccentricities should then evaluate the angular and the module error separately because, as observed here, the performance in one dimension may not go in the same direction as in others.

One limitation of this model is that I decided to change the connectivity parameters of the model based on the observed loss of precision in module error as you get away from fixation (Figure 36A), but other

parameters of the model can also account for broader bumps (i.e. strength of conductance parameters, different connectivities for external inputs, etc.). Another limitation is that I did not correct for cortical magnification (Daniel & Whitteridge, 1961; J. Rovamo & Virsu, 1979) nor for presentation times with eccentricity (Carrasco et al., 2003), so I cannot discard the stimulus size and the speed information processing to play a role in the observed results.

As observed in this chapter, interference effects are a great tool to evaluate the efficiency of computational models because, despite their small magnitude (few degrees), they provide a lot of insight about the topography and the internal circuitry of the network. For instance, without a ring structure, it is not clear how to account for both attractive and repulsive effects. A model explaining changes in behavior with eccentricity can be extremely useful to understand intra- subjects variability (different individual tuning for different eccentricities) and experiments observing contradicting results (Bae & Luck, 2017; Rademaker et al., 2015). Furthermore, this chapter could make the field consider interference effects not only as distance-between-stimuli effect but as a combination of both eccentricity and distance-between-stimuli. Moreover, this chapter will contribute to orient the field towards a more circuitry-based debate, so we all propose specific models and mechanisms by which the biases found in behavior are explained.

In this chapter, I separated perceptual from memory effects in the same vsWM paradigm, so I could evaluate the validity of the SRT (Christophel et al., 2017; Gayet et al., 2018; Scimeca et al., 2018; Xu, 2018) by indirectly addressing where is WM likely to be maintained. Thanks to this controlled paradigm, and by using a measure of error that cleaned the topographical systematic errors (*Methods-Paradigms and analysis*), I found repulsive effects for the perceptual condition and both attractive and repulsive effects for the WM condition (Figure 37). This difference in the direction of the effects pointed towards different networks being responsible for the encoding and WM maintenance. The SRT would predict the same biases for perception and memory, as they rely in the same neural substrate. Results are consistent with Harrison & Bays (2018), where they used critical spacing of visual crowding to evaluate how the WM content was affected by overlapping representations. While the SRT would predict an impairment, as they would be stimulating the same cortical space, they

did not observe an effect of cortical crowding in performance. From my point of view, the study of Harrison & Bays (2018) had two limitations that are solved in this thesis: first, they made a conclusion out of a negative behavioral result (no differences in performance was observed) and second, as stimuli were presented sequentially, it is still possible that the WM content had enough time to shift to neurons that normally encode sensory stimulation in some other part of the visual field (Ester et al., 2009). Both issues are solved here, and on top, a computational model that mechanistically explains the interference is provided.

According to the bump attractor computational model, attractive and repulsive interference effects depend on the neural tuning (Almeida et al., 2015; Nassar et al., 2018). Tuning in sensory areas is shown to be quite stable over time (Bachatene et al., 2015; Lütcke et al., 2013), so different tuning-based results for encoding and WM already point to these processes taking place in different areas. Although attention can sharp tuning properties (Compte & Wang, 2006; David et al., 2008; McAdams & Maunsell, 2000), the increase of tunning would lead to a change from the attractive to the repulsive regime and not the other ways around, which is the observed pattern (Figure 37). The observed results and a potential sharpening of the tuning curves through attentional modulation would support the evolution-inspired idea that sensory areas present repulsive effects to increase visual discrimination while higher order regions with broader tunings present attractive effects to get a more efficient maintenance (Chunharas et al., 2019; Fritsche et al., 2017). Overall, this chapter shows it is necessary to include a no-memory condition (delay 0) to separate perception-based and memory-based effects, so we do not incorrectly attribute to memory, biases that can be due to perceptual processes (Schutte & DeGirolamo, 2020). Besides, it provides convincing evidence against the SRT based on the topographical predictions of it.

Neurophysiological data is essential to validate models. In this case, the bump attractor computational model predicted a loss of tuning with eccentricity. Critically, when checking the tuning curves of PFC in macaques performing a similar task (*Methods-Electrophysiology*, Dataset 1), I observed such loss of tuning with eccentricity (Figure 40). Previous works addressing performance with eccentricity for low-level visual functions (Cowey & Rolls, 1974; Jyrki Rovamo et al., 1978) detected a decrease in performance that could be explained by an increase of the

receptive fields sizes with eccentricity (Johnston & Wright, 1986). Electrophysiological inspections of the macaques visual cortices (Burkhalter & Van Essen, 1986; Desimone & Schein, 1987; Felleman & Van Essen, 1987; Gattass et al., 1981; Toet & Levi, 1992) and also neuroimaging studies in humans (Harvey & Dumoulin, 2011) reported a proportional increase of receptive fields sizes with eccentricity. This linear scaling factor implies the tuning of the neurons of the visual cortex to be constant with eccentricity, inconsistent with the modeling explanation proposed here. Instead, PFC recordings showed the loss of tuning predicted by the computational model. Some studies questioned the proportional increase of receptive fields with eccentricity in visual cortices (Dow et al., 1981; Van Essen et al., 1984) and reported a flatter regime for close eccentricities and a stepper regime for further eccentricities. As I did not have access to any other neural recordings rather than PFC performing this task, I cannot totally discard two different populations in the visual cortex with different topographies nor other regions reproducing the topography observed here, making PFC redundant. In sum, I propose a model that accounts for memory topographical effects both for single-item and multiple-item trials. This model receives already biased information from a perceptual circuit, which cannot overlap with the one responsible for WM due to the nature of the observed interference patterns.

In the second section of the chapter (*Radial dimension*), I developed a radial version of the bump attractor model that gives a mechanistic explanation to a delay-dependent behavioral bias: the time evolution of the compression of the visual space towards the fixation point (Figure 42). Sheth & Shimojo (2001) discussed a possible mechanism based on lateral connections where a constantly activated representation of the fixation point would attract the representations of more eccentric items. However, their interpretation had two important inconsistencies: first, it needed a constant activation of the reference (fixation point). This need was incompatible with another of their results: the compression of the visual space was still present when the referent point was invisible. Second, it needed a transient activity of the target, which does not match with electrophysiological data pointing towards persistent activity of the target during the delay period (Funahashi et al., 1989; Fuster & Alexander, 1971; Goldman-Rakic et al., 1990; Kubota & Niki, 1971). The proposed model for the radial dimension solves both inconsistencies by relying on the topography of the network (the model did not require a bump at the

fixation point to obtain the compressive effect), and on persistent activity (the activity of the target is not restricted to the presentation period, and it is maintained throughout the delay period). This model is the first implementation of the bump attractor model to the radial dimension. In the first section of this chapter, I explained behavioral results through connectivity changes at different eccentricities. In this section, the changes in eccentricity were also implemented by compromising the tuning of more eccentric neurons (Figure 42A) and, as before, I obtained broader bumps for more eccentric locations (Figure 42D-F). In the angular formulation of the bump attractor model, broader bumps diffused less during the delay period as they absorbed noise fluctuations more efficiently. However, in this model formulation of the radial dimension, more eccentric bumps drift more as the pattern of connectivity with neighbor neurons is not constant. Altogether, the angular version and the radial version of the bump attractor model account for a larger error with eccentricity in the radial dimension and a smaller error with eccentricity in the angular dimension. These results open the door to future modeling work that aim to unify both models to explain the error in Euclidean space.

One important difference between behavior and the model is the delay 0 condition (Figure 41C, grey and Figure 42B, light blue). The model does not show any effect (flat line at 0) while behavior shows that the compression of the visual space has a perceptual component, in line with Sheth & Shimojo (2001). As I previously discussed in the angular model, the bump attractor model just explains the delay-dependent effects (not perceptual biases), so different models of visual areas should be considered to explain the biases in early visual cortices observed in both sections (Chunharas et al., 2019; Fritsche et al., 2017; Gibson & Radner, 1937; O'Toole & Wenderoth, 1977). This chapter explained results through two different circuits: one for the encoding of visual information and another for WM maintenance. Future studies may consider evaluating it by using groups of patients with one of this circuits impaired. Patients with optic ataxia, which is a purely perceptual deficit, would be a good candidate to test this hypothesis, as two independent networks would predict no interaction between delay and group (patient vs control) in the compression of the visual space. In other words, the model predicts the same slope in the evolution of the compression of the visual space with delay for optic ataxia patients and controls, although the first ones would present higher

baseline levels of compression. In sum, the second section of this chapter presents the first radial version of the bump attractor computational model, which successfully models the delay component of the compression of the visual field.

In each section of the chapter (*Angular dimension* and *Radial dimension*), I developed a model that was compatible with the one of the other section, as I compromised tuning with eccentricity: κ in the network model of the angular dimension (Equation 3) and S(λ) in the network model in the radial dimension (Equation 9). The fact that equivalent modifications explained behavior in both dimensions demonstrates the flexibility of the bump attractor model. Future experiments must be considered to elucidate if a combined model is needed, because it could be the case that the angular and the radial component of the final memory readout are computed independently in the brain. Alternatively, if the final memory readout was a xy combination in Euclidean space, a two-dimensional model that integrates the angular and the radial dimension would be needed. In the former scenario, the two models presented here could be a good starting point, as they share mechanisms for WM maintenance (PA) and use equivalent parameters to describe the connectivity between neurons.

Distractor filtering in the working memory circuit

In this chapter, I manipulated distractors both in the temporal and the similarity domains. Regarding the temporal domain, changes in TDOA were in line with previous studies (Figure 44, Figure 45, Figure 47 and Figure 48) showing more interference for short TDOAs (Jolicœur & Dell'Acqua, 1998; Pasternak & Zaksas, 2003; Suzuki & Gottlieb, 2013; Van Ede et al., 2018; Vogel et al., 2006). Results suggest that memories go through a "stability process" after encoding (Barbosa, 2017), which may seem in contradiction with another consistent finding of errors increasing with delay length (Funahashi et al., 1989). Although I did not directly evaluate this contradicting effects by playing with very short delays, previous works indeed found a decrease of performance for very short delay periods of 400ms compared to delay periods of 1s (Loft et al., 2014; Loft & Remington, 2013). The computational model presented in this chapter would explain both effects through two different sources of errors: encoding-instability and memory-diffusion. The firing-rate dip at early stages of the delay period (Figure 14) explains the encoding-instability related errors, while memory diffusion due to noise fluctuations explains the increased error with delay length (Figure 11).

Regarding the similarity domain, I evaluated the effect of different distances between stimuli (Figure 46). Results were consistent with previous studies, showing attractive interference for close stimuli and repulsive interference for distant stimuli (Almeida et al., 2015; Nassar et al., 2018). Although the observed pattern of interference provides supporting evidence for the ring connectivity of bump attractor model and, there are some intriguing results when comparing the distance effects with the ones of chapter 1 (*Topography of the working memory circuit*). In chapter 1, the interference was between two targets while in this chapter, the interference was between a target and a distractor. Observing that the pattern of interference follows the same topographical rule (attraction for close and repulsion for far), points to all interferences taking place in the same circuit, instead of some areas just maintaining unbiased target information (Iamshchinina et al., 2021; Lorenc et al., 2018). The bump attractor model supports this interpretation in PFC, being it responsible for the final memory readout.

Altogether, the computational model explained behavioral results (TDOA and distance effects, Figure 62 and Figure 63) through incorporating STP mechanisms. Previous studies combined PA with STP mechanisms to explain serial biases (Barbosa et al., 2020; Kilpatrick, 2018; Stein et al., 2020), diffusion (Itskov et al., 2011), or even attractive effects of close distractors (Seeholzer et al., 2019). The motivation for this combinations of mechanisms originated in previous studies showing memory reactivation for unattended stimulus after unspecific stimulation of the circuit (Rose et al., 2016; Wolff et al., 2017), which planted the idea that other mechanisms not relying on PA were needed, such as synaptic mechanisms (Fiebig & Lansner, 2017; Mongillo et al., 2008). Although posterior experiments showed that those results could be a consequence insufficient statistical power to detect PA (Barbosa et al., 2021; Christophel et al., 2018), memory mediated by PA alone is insufficient to explain some electrophysiological findings. Barbosa et al. (2020) observed that after response, decoding of the target reduced to chance levels during the ITI. However, decoding reappeared specifically during the fixation period of the next trial, in the direction of the observed serial. A version of the bump attractor model without STP would explain behavioral results through a bump of activity staying during the ITI and interfering with the next trial. The latter explanation, however, would be inconsistent with chance level decoding during the ITI, so Barbosa et al. (2020) needed to incorporate STP mechanism. The model presented here needed STP to explain short vs long TDOA effects as well as interference from previously presented distractors. In the model, the elevated firing during the stimulus presentation induces an initial STD that takes time to recover, creating a window of instability that explains why short-TDOA conditions are more disruptive than long-TDOA conditions. Also, distractors are not maintained in the form of PA in the order 2 condition (Figure 62 and Figure 63), so their interference is mediated by their synaptic trace, which is also more degraded for long TDOAs. As in Barbosa et al. (2020), these results would not be explained with the original bump attractor model without STP (Compte et al., 2000), as it would not have an instability period following stimulus presentation and previously presented distractor would not be attenuated.

The dip in the firing rate proposed by the model has been observed in the classical Funahashi et al. (1989) when aligning to the stimulus presentation (Figure 5B), and the electrophysiological analysis of the firing rates of the

Dataset2 (Suzuki & Gottlieb, 2013) showed an interesting correlation between a delay-dip and the final decoding (Figure 67C). Although there is certain circularity in the analysis as distractors were presented during the delay period and because neurons for the original paper were selected depending on their cue selectivity, a null correlation would indicate that those neurons responsible for the final memory readout are the ones that present a stable code with no instability during the delay. Previous studies using cross-temporal decoders with distractors showed transient instability when distractors were presented (Hallenbeck et al., 2021; Parthasarathy et al., 2017), consistently with the data in Figure 64A. Previous studies also identified two different periods of stable generalization: one for the stimulus presentation and one for the delay period (Mendoza-Halliday & Martinez-Trujillo, 2017; Parthasarathy et al., 2017; Spaak et al., 2017; Stokes et al., 2013). I observed a consistent asymmetry in the cross-temporal decoding matrix: while training during the stimulus presentation did not allow to decode during the delay period, training during the delay period allowed to decode during stimulus presentation (Figure 64A, Figure 66B). One plausible explanation for it relies on the procedure of selecting neurons, as neural heterogeneity with PFC subpopulations with selectivity to different time intervals is been reported (Markowitz et al., 2015; Mendoza-Halliday & Martinez-Trujillo, 2017). Both Markowitz et al. (2015) and Mendoza-Halliday & Martinez-Trujillo (2017) observed neurons with selectivity to the stimulus presentation, neurons with selectivity to the delay period and neurons with selectivity to both. As neurons included in the cross-temporal decoding analysis were selected based on their stimulus selectivity, one of the three previously described subpopulations -selectivity to the delay period- is not represented. Therefore, neurons with delay-selectivity would also have stimulus-selectivity, generating the asymmetry. If all the different types of neurons could be incorporated into the analysis, it is possible this asymmetry disappears. Markowitz et al. (2015) already pointed out that the subpopulation with both delay and visual stimulus selectivity was the one with better correlation with behavior, so a future analysis comparing the correlation of the slope of the delay-selectivity subpopulation with decoding strength is likely to reveal a worse correlation than the one presented here. Markowitz et al. (2015) did not observe a dip in the rate at the early delay of the subpopulation with selectivity to both stimulus and delay, so an alternative model to the one

presented here would rely in the communication between the neurons with delay-selectivity and the ones with both. In this alternative model, TDOA effects would probably be explained through the population with delay selectivity, which shows a ramping effect, in a similar way as in the model of Murray, Jaramillo et al. (2017) but by changing the parietal cortex by another subpopulation. This alternative model, however, would have some inconveniences. First, it would be less parsimonious, as it would require more populations to explain behavior; second, some of this subpopulations have not been found in similar experiments (Tsujimoto & Sawaguchi, 2004); third, when found, these subpopulations were detected at different locations of the PFC (Markowitz et al., 2015), which could potentially affect their communication; and fourth, to explain behavioral effects -such as serial biases- this model would still require STP mechanisms. All things considered, future modeling effort should be put into balancing PA, STP and different subpopulations to explain both electrophysiology and behavior.

While in the first chapter of the results (*Topography of the working memory circuit*) I used the bump attractor model exclusively to explain behavioral results, in this chapter I also used it to make predictions. The bump attractor model explains interference effects through the coexistence of the target and the distractor in the same network under similar codes. Therefore, a model trained to decode the target should be able to decode the distractor and vice versa. To test this, I tried to reconstruct WM content in visual, parietal, and frontal areas training both in the delay period of the target and in the delay period of the distractor. When training in the delay period of the target, I could systematically reconstruct the WM content of the target in all three regions (Figure 50A, Figure 55, red), as expected due to the distributed nature of WM (Christophel et al., 2017) and previous studies using IEM to decode information (Ester et al., 2013; Hallenbeck et al., 2021; S. A. Harrison & Tong, 2009; Lorenc et al., 2018; Rademaker et al., 2015, 2019; Serences, 2016; Serences et al., 2009). Critically, when training in the distractor delay period, decoding the target was only possible in frontal areas (Figure 50B, Figure 55, grey), in line with the prediction of the model. Not being able to decode the target when training in the distractor delay period in visual and parietal could be interpreted in two different ways: first, visual and parietal areas have a mnemonic code that is sensory-dependent (Rademaker et al., 2019; Sprague et al., 2014, 2016), so they need a model

trained in the best possible attentional conditions. This first interpretation would predict that, when training on the target delay period, the distractor should be decoded. On the other hand, visual and parietal areas could have two different and independent codes, one for the target and one for the distractor, in line with recent studies that showed sensory information remapped or rotated into subspaces to avoid interference (Libby & Buschman, 2021; Wan et al., 2020), which is somehow represented in the patterns of BOLD signal. This second interpretation would predict that, when training on the distractor delay period, the distractor should be decoded. When tested this hypothesis, I observed that decoding the distractor was possible when training in the target (Figure 58), in line with previous results using IEMs. In frontal regions, decoding the distractor with either training in the or the distractor gave similar results, which is the predicted result for them coexisting in the same circuit.

In this thesis there are several differences to discuss compared to previous studies using IEM to reconstruct the WM content. First of all, previous studies used an independent task to train the IEMs (Ester et al., 2015; Sprague et al., 2014, 2016) while others used the actual WM task (Hallenbeck et al., 2021; Lorenc et al., 2018; Rademaker et al., 2019). Training the IEM in the independent task (*Methods-MRI*) just allowed me to reconstruct during the stimulus presentation in visual (results not included), while using the delay period of the actual WM task worked better. In the very influential paper Rademaker et al. (2019), they observed that training in the independent task allowed them to decode orientations in visual areas, but for parietal areas they needed to train the model in the delay period of the WM task. As both Rademaker et al. (201) and Iamischinia et al. (2021) pointed out, training in the independent task (with flickering stimulation during the delay period) associates the mnemonic code to sensory-like stimulation and, consequently, if the mnemonic code differs from that configuration, as it is likely to happen as the retinotopic structure is lost, training in the actual delay period with cross-validation procedures is a better strategy to reconstruct the WM content, as this approach capitalizes on any signal differentiating the WM content during the delay. Although my results go in that direction, the fact that I could not decode during the delay period using the independent task even in visual area suggests that, in my dataset, the main problem was the difference between the delay period of the independent task and the WM task. Both the number of stimuli (1 vs 6) and the task difficulty (understand

a cue, increased load, distractors, precise response) suggest that even in visual areas, the pattern of BOLD signal associated with WM is changed. Previous studies have showed how top-down signals can affect sensory areas (Ardid et al., 2007, 2010; Gregoriou et al., 2014; Posner & Gilbert, 1999; Reynolds et al., 2000; Reynolds & Chelazzi, 2004; Treue & Maunsell, 1996), providing a reasonable explanation for obtaining decoding just when training the IEM in the delay period of the WM task even for visual areas.

Regarding the ROI definition, it is important to note several differences compared to previous studies. First, I was not able to retinotopically define as many regions as other studies with that amount of spatial resolution. In visual cortex, I analyzed the activity in V1 and not in other regions like V2, V3, V3AB or V4. As I just explored the most primary area of visual cortex, I must be cautious regarding my interpretations, as other areas may be maintaining the WM content differently. Previous results performing a deeper exploration did not find relevant differences in the role of these areas when applying IEMs. Hallenbeck et al. (2021) combined V1-V3 activity and found differences supporting the maintenance in WM content in these regions compared to V3AB or V4, while Rademaker et al. (2019) could decode systematically with equivalent results from V1 to V4. In parietal regions, I could not systematically separate from subject to subject the areas of intraparietal sulcus (IPS): IPS0, IPS1, IPS2 and IPS3 as previous studies did (Hallenbeck et al., 2021; Rademaker et al., 2019; Sprague et al., 2014, 2016). Instead, I combined the WM localizer mapping with an atlas to get parietal activity associated to WM. While the above mentioned studies found differences between visual and parietal, they just observed a small difference between IPS0 and IPS1 (Rademaker et al., 2019), being the first one more similar to visual cortex. The decoding results of the thesis showed similar patterns of decoding between visual and parietal areas, which could also be explained by the ROI definition of parietal incorporating sensory-related IPS regions. Finally, compared to Hallenbeck et al. (2021), Sprague et al. (2014) or Sprague et al. (2016), I did not systematically find activation in the superior precentral sulcus (sPCS), but did in other areas of superior frontal (Figure 23). A final important limitation in the interpretations of the results is that the electrophysiological results in the monkey brain that motivated the bump attractor model came from PFC, which equivalent location in the human brain would be in more anterior areas than the ROI defined. As several

recording in different frontal areas have shown delay-dependent elevated activity during the delay period (Leavitt et al., 2017), I speculate the proposed attractor dynamics to be generalizable in frontal areas and not specifically to PFC.

It has been proposed that the task demands are essential to discriminate the role of each region in vsWM (Bettencourt & Xu, 2016; Christophel et al., 2017; Lorenc et al., 2018, 2021) and that, consequently, the idea of a singular essential storage could limit our comprehension of WM (Iamshchinina et al., 2021). In this line, previous electrophysiological and neuroimaging data suggest that tasks requiring the memorization of detailed low-level visual features would engage sustained mnemonic activation in early visual areas, while tasks involving the storage of abstract concepts would recruit association areas. Lorenc et al. (2018) observed consistent results with this hypothesis because, in the absence of distractors, the reconstruction of the WM content was more accurate in visual areas. However, when distractors were presented, most accurate reconstructions were observed in parietal regions. In the same direction, Iamischinia et al. (2021) reanalyzed Rademaker et al. (2019) and observed that parietal areas were more resistant against distractors. However, both studied showed that visual areas reproduced behavioral results more consistently and following studies like Hallenbeck et al. (2021) observed a clear correlation between the decoded memory error and the behavioral error in visual areas (Figure 7D) but not in downstream regions. As IEMs provide more precise reconstructions in sensory regions due to the retinotopic structure of the areas (Sprague et al., 2016), it is hard to conclude that WM representations are biased just in sensory regions. If that was not the case, maybe visual areas would be receiving a biased top-down input. The results presented here are consistent with this final interpretation, as I observed a cw-ccw interference in all regions, including frontal areas (Figure 60). The interference was stronger in visual areas, but the fact that it was still present in parietal and frontal areas challenges previous works suggesting downstream areas only hold distractor-resistant representations (Lorenc et al., 2018).

Furthermore, the analysis of the decoding of the target and the distractor at different TDOA conditions (Figure 61) revealed that frontal regions had the most consistent decoding with behavior: higher representation for the target in the less distracting condition and higher distractor reconstruction

for the more distracting one. Previous studies suggested distractor filtering was mediated via a mechanisms that increase the fidelity of the remembered information (Fischer & Whitney, 2012), while other supported mechanisms of pure distractor suppression or a combination of both (Bettencourt & Xu, 2016; Lorenc et al., 2018; Suzuki & Gottlieb, 2013). Both scenarios predict a reduced decoding of the target for higher distracting conditions, either for an active reinforcement of the representation of the remembered angles in the low distracting condition or because of a memory loss due to interference in the high distracting condition. Figure 61A showed this effect in all the regions. The decoding of the distractor was crucial, as frontal was the only region presenting a significant difference in decoding between the low and high distracting condition (Figure 61C). Therefore, results suggest that distraction is mediated by both an increased fidelity of the target but also by a suppression of the distractor. Critically, I observed no decoding for the less distracting condition (no significant interference is observed), so a gate mechanism that prevents the distractors enter the circuit is also likely to take place, maybe through the basial ganglia, as previous literature suggests (Awh & Jonides, 2001; Badre, 2012; Chatham & Badre, 2015; Frank et al., 2001; McNab & Klingberg, 2008). However, when the gating mechanism fails, the results of this thesis are consistent with the distractor being represented in frontal areas and interfering with the target in the same circuit, consistent with the bump attractor model.

One possible explanation for the observed interference compared to some previous studies (Iamshchinina et al., 2021; Lorenc et al., 2018; Rademaker et al., 2019) could be that I am using positions instead of gratings. Perceptual effects and after-images are stronger with gratings (Wade et al., 1996), and they also require an extra step of abstraction compared to dots (Hubel & Wiesel, 1968; Lampl et al., 2001), which could affect their bottom-up processing and their top-down modulation. As Hallenbeck et al. (2021) used positions and still found no correlation with behavior in frontal areas, it is possible that, although distractors are used, tasks are still very easy and behavioral effects are so weak that are hardly detected in downstream areas. Taking together the results of this thesis with previous studies analyzing the distributed nature of WM (Christophel et al., 2017; Leavitt et al., 2017), decoding results in sensory and downstream areas (Rademaker et al., 2019; Sprague et al., 2014) and correlations of decoding with behavior in different tasks demands (Hallenbeck et al.,

2021; Iamshchinina et al., 2021; Lorenc et al., 2018), I consider the following as the most plausible explanation for how the final memory readout is computed: 1) frontal regions are in charge for WM maintenance and send top-down signals to high resolution areas for the accurate final response. 2) The final memory readout is computed as a weighted measure of different WM storage buffers and, the weight for each buffer depends on the task demands: easy tasks could weight strongly visual buffers compared to frontal ones, while hard tasks would weight strongly frontal compared to visual ones. The results obtained in this thesis are compatible with both interpretations, and further studies must be run to discriminate between them.

In the last section of the chapter, I evaluated the adequacy of the bump attractor model to mechanistically explain WM improvement after NB stimulation (Qi et al., 2021). Because correct performance depends on a highly tuned and sensitive system, alterations are mostly functionally detrimental (Bisley et al., 2001; Brozoski et al., 1979; Cañas et al., 2018; Croxson et al., 2011; D'Esposito & Postle, 2015; Duan et al., 2015; Maisson et al., 2018; Major et al., 2015; Opris et al., 2005; K. H. Pribram et al., 1964; Rahman et al., 2021; Yue et al., 2021; Y. Zhang et al., 2021). For this reason, behavioral impairments are also harder to interpret: WM may worsen because of induced drowsiness, or a toothache, and not inform us specifically about mechanisms of WM maintenance. On the other hand, WM improvement is more informative regarding the underlying mechanisms and, consequently, more useful for the development of therapeutic strategies for those diseases where WM is affected. Previous studies regarding WM improvement reported it under computerized training programs (Jaeggi et al., 2008; Klingberg et al., 2005; Sattari et al., 2019), optogenetic manipulations of the glutamatergic circuit (Cardoso-Cruz et al., 2019; K. W. Wang et al., 2019), TMS applied to the dorsolateral PFC (Beynel et al., 2019; Brunoni & Vanderhasselt, 2014; Luber et al., 2007; Ramaraju et al., 2020) or pharmacological interventions (Arnsten, 2006; Aultman & Moghaddam, 2001; Bäckman & Nyberg, 2013; Bontempi et al., 2003; Floresco & Jentsch, 2011; Jäkälä et al., 1999; Spinelli et al., 2006). In some cases, providing electrophysiological or neuroimaging traces correlating with it (Constantinidis & Klingberg, 2016; Garavan et al., 2000; Jolles et al., 2010; McNab et al., 2009; Meyer et al., 2011; Meyers et al., 2012; Olesen et al., 2004; Qi & Constantinidis, 2012a, 2012b; Takeuchi et al., 2016; Tang et al., 2019). However, as opposed to the results presented

here, the mechanistic explanation of how these interventions affected the WM circuit was missing.

As in the model presented in *Topography of the WM circuit: Angular dimension*, the model of this section used broader bumps to reproduce behavior: broader bumps diffuse less during the delay period as they absorb noise fluctuations (Figure 68D). The way to get broader bumps, however, was completely different: while in *chapter 1* broader bumps emerged from wider connectivity profiles in eccentricity (changes in κ of the von Misses distribution), in this section broader bumps were a consequence of an increase in the external input (I_0^E): as NB stimulation provokes a release of ACh in PFC, this would result in blockade of hyperpolarizing intrinsic currents in excitatory neurons, so it should be modeled as an increase in the excitability of the population.

Previous works showing improvement in cognitive tasks (Blake et al., 2017; Dasilva et al., 2019; Galvin et al., 2020; Liu et al., 2017, 2018; Sun et al., 2017; Y. Yang et al., 2013) did not show the condition-dependence observed in Qi et al. (2021), where, for prospective distractors located close to the target, NB stimulation produced WM impairment and, for the other conditions, NB stimulation produced WM improvement (Figure 69). The bump attractor model was still able to replicate this singularity in the dataset, as broader bumps under NB stimulation would also be more susceptible to merging when other stimuli are presented close-by (Figure 69A&C). Explaining this singular condition was crucial, as it showed that this mechanism could also have side effects when inspecting different domains of WM. In this line, I think that future interventions regarding WM should us complex behavioral tasks because, as observed here, improving in some conditions may be detrimental for others.

To my knowledge, this is the first work providing a circuit-level model that successfully reproduces experimental results on WM improvement and impairment upon specific task manipulations. Besides, although anticipation effects have been reported in PFC for predictable stimulus timings (Altamura et al., 2010; Berchicci et al., 2015), examples of how this expectation affects memory traces are rare. In this sense, Qi et al. (2021) showed how expectation under NB stimulation generated elevated firing rate at the precise moment in which the first visual stimulus would have been presented (Figure 70). The model interprets this as the spontaneous generation of a spurious memory in the network because of increased

excitability by NB stimulation and by anticipatory timing signals. Although this is similar to the memory reactivation of mnemonic traces imprinted in synaptic modifications (Barbosa et al., 2020; Mongillo et al., 2008), the relation to events in previous trials is not necessary and these "phantom bumps" could reveal false working memories, where the memory appears at a random location unrelated to previous memory traces. Finally, although sharper tuning at the neuronal level is typically associated with increased performance (Li et al., 2004; Raiguel et al., 2006; Sanayei et al., 2018; T. Yang & Maunsell, 2004), the modeling results presented in this thesis (Figure 71 and Figure 72) are in line with other theoretical studies showing that broader tuning curves can produce either worse or better performance at the whole network level depending on specific conditions (Butts & Goldman, 2006; Ma et al., 2006; Pouget et al., 1999; Stein et al., 2020; K. Zhang & Sejnowski, 1999). In sum, Qi et al. (2021) presented intriguing and unexpected results after NB stimulation, and I showed that the bump attractor model can consistently account for them via an increase in neuronal excitability, which is a reasonable alteration of the neocortical circuit after the release of ACh caused by NB stimulation.

An important limitation of the model is that it is composed of one single type of neurons. In Qi et al. (2021), many neurons that had no selectivity or did not present increased activation during the fixation period were discarded. When I increased the complexity of the model by using a spiking network version of the bump attractor model (which mimicked in a more realistic way the heterogeneous tuning and connectivity of real data), I observed a better reproduction of electrophysiological results, so it is likely that an even more complex model that includes more neuron diversity (Finn et al., 2019; Markowitz et al., 2015; Mendoza-Halliday & Martinez-Trujillo, 2017) could reproduce the results more accurately.

The models proposed in the second chapter show different control mechanisms for distractor filtering: while on model (*Methods-Computational modeling*, Network model for distractor filtering) controlled which of the two stimuli presented was the target through changes in the excitability (I_0^E), the other (*Methods-Computational modeling*, Network model of NB stimulation) relied on changes in the strengths of inhibitory and excitatory connections to modulate the network from a *Remember 1st* to a *Remember 2nd* regime. These different strategies are in line with the task design of each paradigm: while the first

model needed to change between *Remember 1st* (*order 1)* and *remember 2nd* (*order 2*) in a trial basis, so a rapid adaptation is needed, the second did not, as blocks of Remember 1st were presented separately from blocks of Remember 2nd, so more stable control mechanisms could be implemented. Supporting the control mechanism mediated by changes in the excitability, previous studies showed attentional modulation in PFC, which is reflected in baseline activity or in ramping effects before the stimulus presentation (Rainer et al., 1998; Suzuki & Gottlieb, 2013). Supporting the control mechanism mediated by neuromodulation, previous studies proposed different mechanism to achieve neuromodulation on a timescale of seconds (Arnsten, 2010; Arnsten et al., 2012; Vijayraghavan et al., 2017). Although in this thesis I have explored these two control mechanisms, others are still possible. For example, maybe sensory inputs of unwanted distractors get suppressed or attenuated in early sensory regions (Nakajima et al., 2019), so when they reach downstream areas they cannot compete with the targets or they do not even reach memory storage centers; or large-scale brain-area interactions mediate how likely it is a certain stimulus to enter the memory state (Sakai et al., 2002).

In sum, the second chapter provided a deep exploration of how the WM circuit deals with distracting information, combining behavioral, neuroimaging, electrophysiological and modeling results, all of them providing converging evidence towards attractor networks mediating WM maintenance in frontal areas.

General discussion

To conclude, I will discuss the results of the thesis regarding the initial goals. In the first chapter, I explored the topography of the WM circuit to **(1**) evaluate if the topographical relationships are maintained through encoding and maintenance periods of WM. Results with variable delay lengths and eccentricities point towards different circuits mediating encoding and maintenance. As one of the fundamental points of the SRT is that encoding and maintenance occur in the same circuit, my results go against this theory. Furthermore, the bump attractor model, which mechanistically explains WM through PA, successfully explained behavioral results through a change of neural tuning that is not observed in visual cortex. Finally, electrophysiological data confirmed one prediction of the model regarding the RF size in PFC, so all behavioral, modeling, and electrophysiological data converged to attractor dynamics in PFC mediating WM maintenance.

The bump attractor model explains behavior in the angular dimension, leaving the whole radial dimension unexplored. Although I partially explored the radial dimension when addressing the first goal (memory errors at different eccentricities), a mechanistic explanation for errors in the radial dimension was missing. To **(2)** topographically describe the WM circuit on this dimension, I developed a radial version of the bump attractor model that successfully explained the delay dependence of the compression of the visual field, where memory reports are systematically biased towards the fixation point.

In the second chapter, I explored how the WM circuit deals with distracting information. To do so, I **(3)** evaluated the effects of manipulating the distractors in the similarity and the temporal domains. Behavioral results were consistent with previous literature, showing that more similar distractors (in both domains) had higher disruptive effects. Again, the bump attractor model **(4)** could reproduce behavioral results, and predictions of the model **(5)** were tested in neuroimaging and electrophysiological results. Neuroimaging results demonstrated the coexistence of the target and the distractor in frontal areas under the same code and showed the most consistent decoding with behavior in frontal areas. In addition, electrophysiological single-unit PFC recordings

were consistent with the profile of activity predicted by the computational model.

Finally, I explored the validity of the model analyzing a dataset that presented WM improvement after NB stimulation. By manipulating the basal conditions of the model, I **(5)** explained gains in performance, providing the first mechanistic explanation for both WM gains and impairment after an intervention. With this model, I also **(4)** explored different control mechanisms for distractor filtering, as this model relied on a neuromodulation of excitatory and inhibitory neurons, consistent with a block design while the first one relied on a modulation of the excitability of the network, consistent with an interleaved-trial design.

Altogether, this thesis supports a WM maintenance system in frontal areas, with evidence both coming from the topographical and distractor-filtering exploration. The thesis provides evidence towards the bump attractor computational model, which has demonstrated its versatility by explaining changes in eccentricity, under different distractor protocols and under cholinergic modulation. Future exploration must be considered to incorporate other WM buffers into the model, such as parietal or visual, as well as to increase the complexity of the model not just by incorporating STP mechanism but also considering neural heterogeneity.

6.Conclusions

1. The effects of memory precision and simultaneous memory interference at different eccentricities can be explained parsimoniously in a bump attractor network model if we assume a loss of angular neural tuning with eccentricity.

2. Electrophysiological results support a loss of tuning with eccentricity in prefrontal cortex neurons, consistent with the modeling prediction.

3. The same eccentricity-dependent loss of angular neural tuning explains in a novel bump attractor model of the radial dimension the delay-dependent compression of the visual field in working memory.

4. This convergent association of degraded angular tuning in the visual periphery with working memory is in contrast with reported preserved peripheral angular tuning in visual perception. This suggests that the neural circuits responsible for sensory encoding and for working memory maintenance are separate and have quantitatively different topographies.

5. The bump attractor model can explain distractor interference in the similarity and temporal domains through the combination of persistent activity and short-term synaptic plasticity mechanisms.

6. Neuroimaging results support distributed working memory buffers in visual, parietal, and frontal areas.

7. Only in frontal areas, decoding of the target and the distractor is independent of training in the target delay period or in the distractor delay period. This can be interpreted as a validation of the computational model assumption that target and distractor information interfere in the same network, represented with the same code. This suggests that this occurs in frontal cortex.

8. Behavioral results are consistent with the observed decoding in frontal areas, suggesting that distractor filtering is mediated both by an attenuation of distracting information in frontal areas as well as by an increased fidelity of the remembered information.

9. The bump attractor model can account for different control strategies to filter distracting information. Two different control mechanisms, one based on excitability control, and another based on neuromodulatory control explained results for block and interleaved-trial designs, respectively.

10. Electrophysiological and behavioral results after stimulation of the Nucleus Basalis of Meynert during spatial working memory can be explained by increased excitability of neurons in a bump attractor network, which results in broader neural tuning and improved working memory precision.

Bibliography

A

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & De Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, *23*(15), 1427–1431. https://doi.org/10.1016/j.cub.2013.05.065

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Almeida, R., Barbosa, J., & Compte, A. (2015). Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. *Journal of Neurophysiology*, *114*(3), 1806–1818. https://doi.org/10.1152/jn.00362.2015

Altamura, M., Goldberg, T. E., Elvevg, B., Holroyd, T., Carver, F. W., Weinberger, D. R., & Coppola, R. (2010). Prefrontal cortex modulation during anticipation of working memory demands as revealed by magnetoencephalography. *International Journal of Biomedical Imaging*, *2010*. https://doi.org/10.1155/2010/840416

Amari, S. ichi. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*(2), 77–87. https://doi.org/10.1007/BF00337259

Amit, D. J. (1995). The Hebbian paradigm reintegrated: Local reverberations as internal representations. *Behavioral and Brain Sciences*, *18*(4), 617–626. https://doi.org/10.1017/S0140525X00040164

Amit, D. J., & Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex*, *7*(3), 237–252. https://doi.org/10.1093/cercor/7.3.237

Arcaro, M. J., McMains, S. A., Singer, B. D., & Kastner, S. (2009). Retinotopic organization of human ventral visual cortex. *Journal of Neuroscience*, *29*(34), 10638–10652.

https://doi.org/10.1523/JNEUROSCI.2807-09.2009

Ardid, S., Wang, X.-J., & Compte, A. (2007). An integrated microcircuit model of attentional processing in the neocortex. *Journal of Neuroscience*, *27*(32), 8486–8495. https://doi.org/10.1523/JNEUROSCI.1145-07.2007

Ardid, S., Wang, X.-J., Gomez-Cabrero, D., & Compte, A. (2010). Reconciling coherent oscillation with modulation of irregular spiking activity in selective attention: Gamma-range synchronization between sensory and executive cortical areas. *Journal of Neuroscience*, *30*(8), 2856–2870. https://doi.org/10.1523/JNEUROSCI.4222-09.2010

Arnsten, A. F. T. (2006). Stimulants: Therapeutic actions in ADHD. *Neuropsychopharmacology*, *31*(11), 2376–2383. https://doi.org/10.1038/sj.npp.1301164

Arnsten, A. F. T. (2010). The use of α-2A adrenergic agonists for the treatment of attention-deficit/hyperactivity disorder. *Expert Review of Neurotherapeutics*, *10*(10), 1595–1605. https://doi.org/10.1586/ern.10.133

Arnsten, A. F. T. (2013). The neurobiology of thought: The groundbreaking discoveries of patricia Goldman-Rakic 1937-2003. *Cerebral Cortex*, *23*(10), 2269–2281. https://doi.org/10.1093/cercor/bht195

Arnsten, A. F. T., Wang, M. J., & Paspalas, C. D. (2012). Neuromodulation of Thought: Flexibilities and Vulnerabilities in Prefrontal Cortical Network Synapses. In *Neuron* (Vol. 76, Issue 1, pp. 223–239). Neuron. https://doi.org/10.1016/j.neuron.2012.08.038

Aultman, J. M., & Moghaddam, B. (2001). Distinct contributions of glutamate and dopamine receptors to temporal aspects of rodent working memory using a clinically relevant task. *Psychopharmacology*, *153*(3), 353–364. https://doi.org/10.1007/s002130000590

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. In *Trends in Cognitive Sciences* (Vol. 5, Issue 3, pp. 119–126). Elsevier Current Trends. https://doi.org/10.1016/S1364-6613(00)01593-X

B

Bachatene, L., Bharmauria, V., Cattan, S., Rouat, J., & Molotchnikoff, S. (2015). Reprogramming of orientation columns in visual cortex: A domino effect. *Scientific Reports*, *5*. https://doi.org/10.1038/srep09436

Bäckman, L., & Nyberg, L. (2013). Dopamine and training-related working-memory improvement. In *Neuroscience and Biobehavioral Reviews* (Vol. 37, Issue 9, pp. 2209–2219). Neurosci Biobehav Rev. https://doi.org/10.1016/j.neubiorev.2013.01.014

Baddeley, A. (2010). Working memory. *Current Biology*, *20*(4). https://doi.org/10.1016/j.cub.2009.12.014

Baddeley, A., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation - Advances in Research and Theory*, *8*(C), 47–89. https://doi.org/10.1016/S0079-7421(08)60452-1

Badre, D. (2012). Opening the gate to working memory. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 109, Issue 49, pp. 19878–19879). Proc Natl Acad Sci U S A. https://doi.org/10.1073/pnas.1216902109

Bae, G. Y., & Luck, S. J. (2017). Interactions between visual working memory representations. *Attention, Perception, and Psychophysics*, *79*(8), 2376–2395. https://doi.org/10.3758/s13414-017-1404-8

Barbosa, J. (2017). Working memories are maintained in a stable code. In *Journal of Neuroscience* (Vol. 37, Issue 35, pp. 8309–8311). J Neurosci. https://doi.org/10.1523/JNEUROSCI.1547-17.2017

Barbosa, J., & Compte, A. (2020). Build-up of serial dependence in color working memory. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-67861-2

Barbosa, J., Lozano-Soldevilla, D., & Compte, A. (2021). Pinging the brain with visual impulses reveals electrically active, not activity-silent, working memories. *PLoS Biology*, *19*(10), e3001436. https://doi.org/10.1371/journal.pbio.3001436

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. *Nature Neuroscience 2020 23:8*, *23*(8), 1016–1024. https://doi.org/10.1038/S41593-020-0644-4

Barrouillet, P., De Paepe, A., & Langerock, N. (2012). Time causes forgetting from working memory. *Psychonomic Bulletin and Review*, *19*(1), 87–92. https://doi.org/10.3758/s13423-011-0192-8

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science (New York, N.Y.)*, *321*(5890), 851–854. https://doi.org/10.1126/SCIENCE.1158023

Berchicci, M., Lucci, G., Spinelli, D., & Di Russo, F. (2015). Stimulus onset predictability modulates proactive action control in a Go/No-go task. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00101

Bettencourt, K. C., & Xu, Y. (2016). Decoding the content of visual short-term memory under distraction in occipital and parietal areas. *Nature Neuroscience*, *19*(1), 150–157. https://doi.org/10.1038/nn.4174

Beynel, L., Davis, S. W., Crowell, C. A., Hilbig, S. A., Lim, W., Nguyen, D., Palmer, H., Brito, A., Peterchev, A. V., Luber, B., Lisanby, S. H., Cabeza, R., & Appelbaum, L. G. (2019). Online repetitive transcranial magnetic stimulation during working memory in younger and older adults: A randomized within-subject comparison. *PLoS ONE*, *14*(3), 1–19. https://doi.org/10.1371/journal.pone.0213707

Bijsterbosch, J., Smith, S. M., & Beckmann, C. F. (2017). *An Introduction to Resting State FMRI Functional Connectivity.* (Oxford: Ox).

Bisley, J. W., Zaksas, D., & Pasternak, T. (2001). Microstimulation of cortical area MT affects performance on a visual working memory task. *Journal of Neurophysiology*, *85*(1), 187–196. https://doi.org/10.1152/jn.2001.85.1.187

Blake, D. T., Terry, A. V., Plagenhoef, M., Constantinidis, C., & Liu, R. (2017). Potential for intermittent stimulation of nucleus basalis of meynert to impact treatment of alzheimer's disease. *Communicative and Integrative Biology*, *10*(5–6). https://doi.org/10.1080/19420889.2017.1389359

Bliss, D. P., & D'Esposito, M. (2017). Synaptic augmentation in a cortical circuit model reproduces serial dependence in visual working memory. *PLoS ONE*, *12*(12). https://doi.org/10.1371/journal.pone.0188927

Bontempi, B., Whelan, K. T., Risbrough, V. B., Lloyd, G. K., & Menzaghi, F. (2003). Cognitive enhancing properties and tolerability of cholinergic agents in mice: A comparative study of nicotine, donepezil, and SIB-1553A, a subtype-selective ligand for nicotinic acetylcholine receptors. *Neuropsychopharmacology*, *28*(7), 1235–1246. https://doi.org/10.1038/sj.npp.1300150

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392. https://doi.org/10.1177/0956797610397956

Brewer, A. A., Liu, J., Wade, A. R., & Wandell, B. A. (2005). Visual field maps and stimulus selectivity in human ventral occipital cortex. *Nature Neuroscience*, *8*(8), 1102–1109. https://doi.org/10.1038/nn1507

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience*, *29*(44), 13992–14003. https://doi.org/10.1523/JNEUROSCI.3577-09.2009

Brozoski, T. J., Brown, R. M., Rosvold, H. E., & Goldman, P. S. (1979). Cognitive deficit caused by regional depletion of dopamine in prefrontal cortex of rhesus monkey. *Science*, *205*(4409), 929–932. https://doi.org/10.1126/science.112679

Brunoni, A. R., & Vanderhasselt, M. A. (2014). Working memory improvement with non-invasive brain stimulation of the dorsolateral prefrontal cortex: A systematic review and meta-analysis. *Brain and Cognition*, *86*(1), 1–9. https://doi.org/10.1016/j.bandc.2014.01.008

Burak, Y., & Fieted, I. R. (2012). Fundamental limits on persistent activity in networks of noisy neurons. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(43), 17645–17650. https://doi.org/10.1073/pnas.1117386109

Burkhalter, A., & Van Essen, D. C. (1986). Processing of color, form and disparity information in visual areas VP and V2 of ventral

extrastriate cortex in the macaque monkey. *Journal of Neuroscience*, *6*(8), 2327–2351. https://doi.org/10.1523/jneurosci.06-08-02327.1986

Busse, L., Katzner, S., Tillmann, C., & Treue, S. (2008). Effects of attention on perceptual direction tuning curves in the human visual system. *Journal of Vision*, *8*(9), 1–13. https://doi.org/10.1167/8.9.2

Butts, D. A., & Goldman, M. S. (2006). Tuning curves, neuronal variability, and sensory coding. *PLoS Biology*, *4*(4), 639–646. https://doi.org/10.1371/journal.pbio.0040092

C

Cai, R. H., Pouget, A., Schlag-Rey, M., & Schlag, J. (1997). Perceived geometrical relationships affected by eye-movement signals. *Nature*, *386*(6625), 601–604. https://doi.org/10.1038/386601a0

Cañas, A., Juncadella, M., Lau, R., Gabarrós, A., & Hernández, M. (2018). Working memory deficits after lesions involving the supplementary motor area. *Frontiers in Psychology*, *9*(MAY). https://doi.org/10.3389/fpsyg.2018.00765

Cardoso-Cruz, H., Paiva, P., Monteiro, C., & Galhardo, V. (2019). Selective optogenetic inhibition of medial prefrontal glutamatergic neurons reverses working memory deficits induced by neuropathic pain. *Pain*, *160*(4), 805–823. https://doi.org/10.1097/j.pain.0000000000001457

Carr, D. B., & Surmeier, D. J. (2007). M1 muscarinic receptor modulation of Kir2 channels enhances temporal summation of excitatory synaptic potentials in prefrontal cortex pyramidal neurons. *Journal of Neurophysiology*, *97*(5), 3432–3438. https://doi.org/10.1152/jn.00828.2006

Carrasco, M., McElreel, B., Denisova, K., & Giordano, A. M. (2003). Speed of visual processing increases with eccentricity. *Nature Neuroscience*, *6*(7), 699–700. https://doi.org/10.1038/nn1079

Carter, E., & 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), i16–i26. https://doi.org/10.1093/cercor/bhm103

Chao, L. L., & Knight, R. T. (1998). Contribution of human prefrontal cortex to delay performance. *Journal of Cognitive Neuroscience*, *10*(2), 167–177. https://doi.org/10.1162/089892998562636

Chatham, C. H., & Badre, D. (2015). Multiple gates on working memory. In *Current Opinion in Behavioral Sciences* (Vol. 1, pp. 23–31). Elsevier. https://doi.org/10.1016/j.cobeha.2014.08.001

Chelazzi, L., Miller, E. K., Duncan, J., & Desimone, R. (2001). Responses of neurons in macaque area V4 during memory-guided visual search. *Cerebral Cortex*, *11*(8), 761–772. https://doi.org/10.1093/cercor/11.8.761

Chieffi, S., & Allport, D. A. (1997). Independent coding of target distance and direction in visuo-spatial working memory. *Psychological Research*, *60*(4), 244–250. https://doi.org/10.1007/BF00419409

Christophel, T. B., Hebart, M. N., & Haynes, J. D. (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. *Journal of Neuroscience*, *32*(38), 12983–12989. https://doi.org/10.1523/JNEUROSCI.0184-12.2012

Christophel, T. B., Iamshchinina, P., Yan, C., Allefeld, C., & Haynes, J. D. (2018). Cortical specialization for attended versus unattended working memory. *Nature Neuroscience*, *21*(4), 494–496. https://doi.org/10.1038/s41593-018-0094-4

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J. D. (2017). The Distributed Nature of Working Memory. *Trends in Cognitive Sciences*, *21*(2), 111–124. https://doi.org/10.1016/j.tics.2016.12.007

Chunharas, C., Rademaker, R., Brady, T., & Serences, J. (2019). *Adaptive memory distortion in visual working memory*. https://doi.org/10.31234/osf.io/e3m5a

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X.-J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, *10*(9), 910–923. https://doi.org/10.1093/cercor/10.9.910

Compte, A., & Wang, X.-J. (2006). Tuning curve shift by attention modulation in cortical neurons: A computational study of its mechanisms. *Cerebral Cortex*, *16*(6), 761–778. https://doi.org/10.1093/cercor/bhj021

Constantinidis, C., & Klingberg, T. (2016). The neuroscience of working memory capacity and training. *Nature Reviews Neuroscience*, *17*(7), 438–449. https://doi.org/10.1038/nrn.2016.43

Conway, A. R. A., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin and Review*, *8*(2), 331–335. https://doi.org/10.3758/BF03196169

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Cowey, A., & Rolls, E. T. (1974). Human cortical magnification factor and its relation to visual acuity. *Experimental Brain Research*, *21*(5), 447–454. https://doi.org/10.1007/BF00237163

Croxson, P. L., Kyriazis, D. A., & Baxter, M. G. (2011). Cholinergic modulation of a specific memory function of prefrontal cortex. *Nature Neuroscience*, *14*(12), 1510–1512. https://doi.org/10.1038/nn.2971

D

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology*, *66*, 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

Daniel, P. M., & Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys. *The Journal of Physiology*, *159*(2), 203–221. https://doi.org/10.1113/jphysiol.1961.sp006803

Dasilva, M., Brandt, C., Gotthardt, S., Gieselmann, M. A., Distler, C., & Thiele, A. (2019). Cell class-specific modulation of attentional signals by acetylcholine in macaque frontal eye field. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(40), 20180–20189. https://doi.org/10.1073/pnas.1905413116

David, S. V., Hayden, B. Y., Mazer, J. A., & Gallant, J. L. (2008). Attention to Stimulus Features Shifts Spectral Tuning of V4 Neurons during Natural Vision. *Neuron*, *59*(3), 509–521. https://doi.org/10.1016/j.neuron.2008.07.001

De Fockert, J. W., Rees, G., Frith, C. D., & Lavie, N. (2001). The role of working memory in visual selective attention. *Science*, *291*(5509), 1803–1806. https://doi.org/10.1126/science.1056496

Denker, M., Yegenoglu, A., & Grün, S. (2018). Collaborative HPC-enabled workflows on the HBP Collaboratory using the Elephant framework. In *Neuroinformatics* (p. 19). https://doi.org/10.12751/incf.ni2018.0019

Desimone, R., & Schein, S. J. (1987). Visual properties of neurons in area V4 of the macaque: Sensitivity to stimulus form. *Journal of Neurophysiology*, *57*(3), 835–868. https://doi.org/10.1152/jn.1987.57.3.835

Dougherty, R. F., Koch, V. M., Brewer, A. A., Fischer, B., Modersitzki, J., & Wandell, B. A. (2003). Visual field representations and locations of visual areas v1/2/3 in human visual cortex. *Journal of Vision*, *3*(10), 586–598. https://doi.org/10.1167/3.10.1

Dow, B. M., Snyder, A. Z., Vautin, R. G., & Bauer, R. (1981). Magnification factor and receptive field size in foveal striate cortex of the monkey. *Experimental Brain Research*, *44*(2), 213–228. https://doi.org/10.1007/BF00237343

Duan, A. R., Varela, C., Zhang, Y., Shen, Y., Xiong, L., Wilson, M. A., & Lisman, J. (2015). Delta frequency optogenetic stimulation of the thalamic nucleus reuniens is sufficient to produce working memory deficits: Relevance to schizophrenia. In *Biological Psychiatry* (Vol. 77, Issue 12, pp. 1098–1107). Biol Psychiatry. https://doi.org/10.1016/j.biopsych.2015.01.020

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational Models of Working Memory. *Nature Neuroscience*, *3*(11s), 1184–1191. https://doi.org/10.1038/81460

E

Edin, F., Klingberg, T., Johansson, P., McNab, F., Tegnér, J., & Compte, A. (2009). Mechanism for top-down control of working memory capacity. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), 6802–6807. https://doi.org/10.1073/pnas.0901894106

Elmore, L. C., Ji Ma, W., Magnotti, J. F., Leising, K. J., Passaro, A. D., Katz,

J. S., & Wright, A. A. (2011). Visual short-term memory compared in rhesus monkeys and humans. *Current Biology*, *21*(11), 975–979. https://doi.org/10.1016/j.cub.2011.04.031

Emrich, S. M., Riggall, A. C., La Rocque, J. J., & Postle, B. R. (2013). Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. *Journal of Neuroscience*, *33*(15), 6516–6523. https://doi.org/10.1523/JNEUROSCI.5732-12.2013

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E. J., & Shadlen, M. N. (1994). FMRI of human visual cortex [5]. In *Nature* (Vol. 369, Issue 6481, p. 525). Nature Publishing Group. https://doi.org/10.1038/369525a0

Ester, E. F., Anderson, D. E., Serences, J. T., & Awh, E. (2013). A neural measure of precision in visual working memory. *Journal of Cognitive Neuroscience*, *25*(5), 754–761. https://doi.org/10.1162/jocn_a_00357

Ester, E. F., Serences, J. T., & Awh, E. (2009). Spatially global representations in human primary visual cortex during working memory maintenance. *Journal of Neuroscience*, *29*(48), 15258–15265. https://doi.org/10.1523/JNEUROSCI.4388-09.2009

Ester, E. F., Sprague, T. C., & Serences, J. T. (2015). Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. *Neuron*, *87*(4), 893–905. https://doi.org/10.1016/j.neuron.2015.07.013

F

Felleman, D. J., & Van Essen, D. C. (1987). Receptive field properties of neurons in area V3 of macaque monkey extrastriate cortex. *Journal of Neurophysiology*, *57*(4), 889–920. https://doi.org/10.1152/jn.1987.57.4.889

Fiebig, X. F., & Lansner, X. A. (2017). A Spiking Working Memory Model Based on Hebbian Short-Term Potentiation. *The Journal of Neuroscience*, *37*(1), 83–96. https://doi.org/10.1523/JNEUROSCI.1989-16.2016

Finn, E. S., Huber, L., Jangraw, D. C., Molfese, P. J., & Bandettini, P. A. (2019). Layer-dependent activity in human prefrontal cortex during

working memory. *Nature Neuroscience*, *22*(10), 1687–1695. https://doi.org/10.1038/s41593-019-0487-z

Fischer, J., & Whitney, D. (2012). Attention gates visual coding in the human pulvinar. *Nature Communications*, *3*(1), 1–9. https://doi.org/10.1038/ncomms2054

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*, *17*(5), 738–743. https://doi.org/10.1038/nn.3689

Floresco, S. B., & Jentsch, J. D. (2011). Pharmacological enhancement of memory and executive functioning in laboratory animals. *Neuropsychopharmacology*, *36*(1), 227–250. https://doi.org/10.1038/npp.2010.158

Fougnie, D. (2008). The relationship between attention and working memory. In N. B. Johansen (Ed.), *New research on short-term memory*. Nova Science. https://doi.org/10.3389/conf.fnhum.2011.207.00576

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective and Behavioral Neuroscience*, *1*(2), 137–160. https://doi.org/10.3758/CABN.1.2.137

Frick, R. W. (1988). Issues of representation and limited capacity in the visuospatial sketchpad. *British Journal of Psychology*, *79*(3), 289–308. https://doi.org/10.1111/j.2044-8295.1988.tb02289.x

Fritsche, M., Mostert, P., & de Lange, F. P. (2017). Opposite Effects of Recent History on Perception and Decision. *Current Biology*, *27*(4), 590–595. https://doi.org/10.1016/j.cub.2017.01.006

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61*(2), 331–349. https://doi.org/10.1152/jn.1989.61.2.331

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1991). Neuronal activity related to saccadic eye movements in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *65*(6), 1464–1483. https://doi.org/10.1152/jn.1991.65.6.1464

Funahashi, S., Chafee, M. V., & Goldman-Rakic, P. S. (1993). Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task. *Nature*, *365*(6448), 753–756. https://doi.org/10.1038/365753a0

Fuster, J. M. (1988). Prefrontal Cortex. In *Comparative Neuroscience and Neurobiology* (pp. 107–109). Birkhäuser, Boston, MA. https://doi.org/10.1016/B978-008045046-9.01118-9

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, *173*(3997), 652–654. https://doi.org/10.1126/science.173.3997.652

G

Galvin, V. C., Yang, S. T., Paspalas, C. D., Yang, Y., Jin, L. E., Datta, D., Morozov, Y. M., Lightbourne, T. C., Lowet, A. S., Rakic, P., Arnsten, A. F. T., & Wang, M. (2020). Muscarinic M1 Receptors Modulate Working Memory Performance and Activity via KCNQ Potassium Channels in the Primate Prefrontal Cortex. *Neuron*, *106*(4), 649-661.e4. https://doi.org/10.1016/j.neuron.2020.02.030

Garavan, H., Kelley, D., Rosen, A., Rao, S. M., & Stein, E. A. (2000). Practice-related functional activation changes in a working memory task. *Microscopy Research and Technique*, *51*(1), 54–63. https://doi.org/10.1002/1097-0029(20001001)51:1<54::AID-JEMT6>3.0.CO;2-J

Gattass, R., Gross, C. G., & Sandell, J. H. (1981). Visual topography of V2 in the macaque. *Journal of Comparative Neurology*, *201*(4), 519–539. https://doi.org/10.1002/cne.902010405

Gayet, S., Guggenmos, M., Christophel, T. B., Haynes, J. D., Paffen, C. L. E., Van Der Stigchel, S., & Sterzer, P. (2017). Visual working memory enhances the neural response to matching visual input. *Journal of Neuroscience*, *37*(28), 6638–6647. https://doi.org/10.1523/JNEUROSCI.3418-16.2017

Gayet, S., Paffen, C. L. E., & Van der Stigchel, S. (2018). Visual Working Memory Storage Recruits Sensory Processing Areas. In *Trends in Cognitive Sciences* (Vol. 22, Issue 3, pp. 189–190). Trends Cogn Sci. https://doi.org/10.1016/j.tics.2017.09.011

Gibson, J. J., & Radner, M. (1937). Adaptation, after-effect and contrast in the perception of tilted lines. *Journal of Experimental Psychology*, *20*(5), 453–467. https://doi.org/10.1037/h0059826

Gignac, G. E., & Weiss, L. G. (2015). Digit Span is (mostly) related linearly to general intelligence: Every extra bit of span counts. *Psychological Assessment*, *27*(4), 1312–1323. https://doi.org/10.1037/pas0000105

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932. https://doi.org/10.1038/nn.2831

Gnadt, J. W., & Andersen, R. A. (1988). Memory related motor planning activity in posterior parietal cortex of macaque. *Experimental Brain Research*, *70*(1), 216–220. https://doi.org/10.1007/BF00271862

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. In *Neuron* (Vol. 14, Issue 3, pp. 477–485). Neuron. https://doi.org/10.1016/0896-6273(95)90304-6

Goldman-Rakic, P. S., Funahashi, S., & Bruce, C. J. (1990). Neocortical memory circuits. *Cold Spring Harbor Symposia on Quantitative Biology*, *55*, 1025–1038. https://doi.org/10.1101/SQB.1990.055.01.097

Goldman-Rakic, P. S. (1987). Circuitry of Primate Prefrontal Cortex and Regulation of Behavior by Representational Memory. In *Comprehensive Physiology* (pp. 373–417). Wiley. https://doi.org/10.1002/cphy.cp010509

Gregoriou, G. G., Rossi, A. F., Ungerleider, L. G., & Desimone, R. (2014). Lesions of prefrontal cortex reduce attentional modulation of neuronal responses and synchrony in V4. *Nature Neuroscience*, *17*(7), 1003–1011. https://doi.org/10.1038/nn.3742

H

Hallenbeck, G. E., Sprague, T. C., Rahmati, M., Sreenivasan, K. K., & Curtis, C. E. (2021). Working memory representations in visual cortex mediate distraction effects. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-24973-1

Hansel, D., & Mato, G. (2013). Short-term plasticity explains irregular persistent activity in working memory tasks. *Journal of Neuroscience*, *33*(1), 133–149. https://doi.org/10.1523/JNEUROSCI.3455-12.2013

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635. https://doi.org/10.1038/nature07832

Harrison, W. J., & Bays, P. M. (2018). Visual working memory is independent of the cortical spacing between memoranda. *Journal of Neuroscience*, *38*(12), 3116–3123. https://doi.org/10.1523/JNEUROSCI.2645-17.2017

Harvey, B. M., & Dumoulin, S. O. (2011). The relationship between cortical magnification factor and population receptive field size in human visual cortex: Constancies in cortical architecture. *Journal of Neuroscience*, *31*(38), 13604–13612. https://doi.org/10.1523/JNEUROSCI.2572-11.2011

Hebb, D. (1949). *The Organization of Behavior.* Wiley.

Hedrick, T., & Waters, J. (2015). Acetylcholine excites neocortical pyramidal neurons via nicotinic receptors. *Journal of Neurophysiology*, *113*(7), 2195–2209. https://doi.org/10.1152/jn.00716.2014

Hembrook-Short, J. R., Mock, V. L., & Briggs, F. (2017). Attentional Modulation of Neuronal Activity Depends on Neuronal Feature Selectivity. *Current Biology*, *27*(13), 1878-1887.e5. https://doi.org/10.1016/j.cub.2017.05.080

Herwig, A., Beisert, M., & 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). https://doi.org/10.1167/10.5.8

Honda, H. (1993). Saccade-contingent displacement of the apparent position of visual stimuli flashed on a dimly illuminated structured background. *Vision Research*, *33*(5–6), 709–716. https://doi.org/10.1016/0042-6989(93)90190-8

Honda, H. (1995). Visual mislocalization produced by a rapid image displacement on the retina: Examination by means of dichoptic presentation of a target and its background scene. *Vision Research*, *35*(21), 3021–3028. https://doi.org/10.1016/0042-6989(95)00108-C

Hubbard, T. L. (1995). Cognitive Representation of Motion: Evidence for Friction and Gravity Analogues. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 241–254. https://doi.org/10.1037/0278-7393.21.1.241

Hubbard, T. L. (1998). Some effects of representational friction, target size, and memory averaging on memory for vertically moving targets. *Canadian Journal of Experimental Psychology*, *52*(1), 44–49. https://doi.org/10.1037/h0087278

Hubbard, T. L., & Ruppel, S. E. (2000). Spatial memory averaging, the landmark attraction effect, and representational gravity. *Psychological Research*, *64*(1), 41–55. https://doi.org/10.1007/s004260000029

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243. https://doi.org/10.1113/jphysiol.1968.sp008455

Huttenlocher, J., Hedges, L. V., Corrigan, B., & Crawford, L. E. (2004). Spatial categories and the estimation of location. *Cognition*, *93*(2), 75–97. https://doi.org/10.1016/j.cognition.2003.10.006

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and Particulars: Prototype Effects in Estimating Spatial Location. *Psychological Review*, *98*(3), 352–376. https://doi.org/10.1037/0033-295X.98.3.352

I

Iamshchinina, P., Christophel, T. B., Gayet, S., & Rademaker, R. L. (2021). Essential considerations for exploring visual working memory storage in the human brain. *Visual Cognition*, *29*(7), 425–436. https://doi.org/10.1080/13506285.2021.1915902

Itskov, V., Hansel, D., & Tsodyks, M. (2011). Short-term facilitation may stabilize parametric working memory trace. *Frontiers in Computational Neuroscience*, *5*. https://doi.org/10.3389/fncom.2011.00040

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. https://doi.org/10.1038/35058500

J

Jacob, S. N., & Nieder, A. (2014). Complementary roles for primate frontal and parietal cortex in guarding working memory from distractor stimuli. *Neuron*, *83*(1), 226–237. https://doi.org/10.1016/j.neuron.2014.05.009

Jacobsen, C. F. (1936). Studies of cerebral function in primates. I. The functions of the frontal association areas in monkeys. *Comparative Psychology Monographs*, *13*, *3*, 1–60.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(19), 6829–6833. https://doi.org/10.1073/pnas.0801268105

Jäkälä, P., Riekkinen, M., Sirviö, J., Koivisto, E., Kejonen, K., Matti, & Riekkinen, P. (1999). Guanfacine, but not clonidine, improves planning and working memory performance in humans. *Neuropsychopharmacology*, *20*(5), 460–470. https://doi.org/10.1016/S0893-133X(98)00127-4

Jastrow, J. (1892). Studies from the University of Wisconsin: On the Judgment of Angles and Positions of Lines. *The American Journal of Psychology*, *5*(2), 214. https://doi.org/10.2307/1410867

Johnson, J. S., Spencer, J. P., Luck, S. J., & Schöner, G. (2009). A dynamic neural field model of visual working memory and change detection: Research article. *Psychological Science*, *20*(5), 568–577. https://doi.org/10.1111/j.1467-9280.2009.02329.x

Johnston, A., & Wright, M. J. (1986). Matching velocity in central and peripheral vision. *Vision Research*, *26*(7), 1099–1109. https://doi.org/10.1016/0042-6989(86)90044-1

Jolicœur, P., & Dell'Acqua, R. (1998). The Demonstration of Short-Term Consolidation. *Cognitive Psychology*, *36*(2), 138–202. https://doi.org/10.1006/cogp.1998.0684

Jolles, D. D., Grol, M. J., Van Buchem, M. A., Rombouts, S. A. R. B., & Crone, E. A. (2010). Practice effects in the brain: Changes in cerebral activation after working memory practice depend on task demands. *NeuroImage*, *52*(2), 658–668. https://doi.org/10.1016/j.neuroimage.2010.04.028

K

Kane, M. J., Gross, G. M., Chun, C. A., Smeekens, B. A., Meier, M. E., Silvia, P. J., & Kwapil, T. R. (2017). For Whom the Mind Wanders, and When, Varies Across Laboratory and Daily-Life Settings. *Psychological Science*, *28*(9), 1271–1289. https://doi.org/10.1177/0956797617706086

Kerzel, D. (2002). Memory for the position of stationary objects: Disentangling foveal bias and memory averaging. *Vision Research*, *42*(2), 159–167. https://doi.org/10.1016/S0042-6989(01)00274-7

Kilpatrick, Z. P. (2018). Synaptic mechanisms of interference in working memory. *Scientific Reports*, *8*(1), 1–20. https://doi.org/10.1038/s41598-018-25958-9

Klein, A., & Tourville, J. (2012). 101 Labeled Brain Images and a Consistent Human Cortical Labeling Protocol. *Frontiers in Neuroscience*, *6*(DEC). https://doi.org/10.3389/fnins.2012.00171

Klingberg, T. (2009). *The Overflowing Brain*. Oxford University Press.

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., Gillberg, C. G., Forssberg, H., & Westerberg, H. (2005). Computerized training of working memory in children with ADHD - A randomized, controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry*, *44*(2), 177–186. https://doi.org/10.1097/00004583-200502000-00010

Konstantinou, N., Beal, E., King, J. R., & Lavie, N. (2014). Working memory load and distraction: Dissociable effects of visual maintenance and cognitive control. *Attention, Perception, and Psychophysics*, *76*(7), 1985–1997. https://doi.org/10.3758/s13414-014-0742-z

Kubota, K., & Niki, H. (1971). Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology*, *34*(3), 337–347. https://doi.org/10.1152/jn.1971.34.3.337

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*(4), 389–433. https://doi.org/10.1016/S0160-2896(05)80012-1

L

Laming, D., & Laming, J. (1992). F. Hegelmaier: On memory for the length of a line. *Psychological Research*, *54*(4), 233–239. https://doi.org/10.1007/BF01358261

Lampl, I., Anderson, J. S., Gillespie, D. C., & Ferster, D. (2001). Prediction of orientation selectivity from receptive field architecture in simple cells of cat visual cortex. *Neuron*, *30*(1), 263–274. https://doi.org/10.1016/S0896-6273(01)00278-1

Lara, A. H., & Wallis, J. D. (2015). The role of prefrontal cortex in working memory: A mini review. *Frontiers in Systems Neuroscience*, *9*(DEC). https://doi.org/10.3389/fnsys.2015.00173

Leavitt, M. L., Mendoza-Halliday, D., & Martinez-Trujillo, J. C. (2017). Sustained Activity Encoding Working Memories: Not Fully Distributed. In *Trends in Neurosciences* (Vol. 40, Issue 6, pp. 328–346). Trends Neurosci. https://doi.org/10.1016/j.tins.2017.04.004

Li, W., Piëch, V., & Gilbert, C. D. (2004). Perceptual learning and top-down influences in primary visual cortex. *Nature Neuroscience*, *7*(6), 651–657. https://doi.org/10.1038/nn1255

Libby, A., & Buschman, T. J. (2021). Rotational dynamics reduce interference between sensory and memory representations. *Nature Neuroscience*, *24*(5), 715–726. https://doi.org/10.1038/s41593-021-00821-9

Lipinski, J., Simmering, V. R., Johnson, J. S., & Spencer, J. P. (2010). The role of experience in location estimation: Target distributions shift location memory biases. *Cognition*, *115*(1), 147–153. https://doi.org/10.1016/j.cognition.2009.12.008

Liu, R., Crawford, J., Callahan, P. M., Terry, A. V., Constantinidis, C., & Blake, D. T. (2017). Intermittent Stimulation of the Nucleus Basalis of Meynert Improves Working Memory in Adult Monkeys. *Current Biology*, *27*(17), 2640-2646.e4. https://doi.org/10.1016/j.cub.2017.07.021

Liu, R., Crawford, J., Callahan, P. M., Terry, A. V., Constantinidis, C., & Blake, D. T. (2018). Intermittent stimulation in the nucleus basalis of meynert improves sustained attention in rhesus monkeys. *Neuropharmacology*, *137*, 202–210. https://doi.org/10.1016/J.NEUROPHARM.2018.04.026

Loft, S., Doyle, K. L., Naar-King, S., Outlaw, A. Y., Nichols, S. L., Weber, E., Casaletto, K. B., & Woods, S. P. (2014). Allowing brief delays in responding improves event-based prospective memory for young adults living with HIV disease. *Journal of Clinical and Experimental Neuropsychology*, *36*(7), 761–772. https://doi.org/10.1080/13803395.2014.942255

Loft, S., & Remington, R. W. (2013). Wait a second: Brief delays in responding reduce focality effects in event-based prospective memory. *Quarterly Journal of Experimental Psychology*, *66*(7), 1432–1447. https://doi.org/10.1080/17470218.2012.750677

Lorenc, E. S., Mallett, R., & Lewis-Peacock, J. A. (2021). Distraction in Visual Working Memory: Resistance is Not Futile. In *Trends in Cognitive Sciences* (Vol. 25, Issue 3, pp. 228–239). Elsevier Current Trends. https://doi.org/10.1016/j.tics.2020.12.004

Lorenc, E. S., Sreenivasan, K. K., Nee, D. E., & Esposito, M. D. (2018). Flexible coding of visual working memory representations during distraction. *The Journal of Neuroscience*, *38*(23), 5267–5276. https://doi.org/10.1523/JNEUROSCI.3061-17.2018

Luber, B., Kinnunen, L. H., Rakitin, B. C., Ellsasser, R., Stern, Y., & Lisanby, S. H. (2007). Facilitation of performance in a working memory task with rTMS stimulation of the precuneus: Frequency- and time-dependent effects. *Brain Research*, *1128*(1), 120–129. https://doi.org/10.1016/j.brainres.2006.10.011

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Lütcke, H., Margolis, D. J., & Helmchen, F. (2013). Steady or changing? Long-term monitoring of neuronal population activity. *Trends in Neurosciences*, *36*(7), 375–384. https://doi.org/10.1016/j.tins.2013.03.008

M

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438. https://doi.org/10.1038/nn1790

MacOveanu, J., Klingberg, T., & Tegnér, J. (2007). Neuronal firing rates account for distractor effects on mnemonic accuracy in a visuo-spatial working memory task. *Biological Cybernetics*, *96*(4), 407–419. https://doi.org/10.1007/s00422-006-0139-8

Maisson, D. J. N., Gemzik, Z. M., & Griffin, A. L. (2018). Optogenetic suppression of the nucleus reuniens selectively impairs encoding during spatial working memory. *Neurobiology of Learning and Memory*, *155*(June), 78–85. https://doi.org/10.1016/j.nlm.2018.06.010

Major, A. J., Vijayraghavan, S., & Everling, S. (2015). Muscarinic attenuation of mnemonic rule representation in macaque dorsolateral prefrontal cortex during a pro-and anti-saccade task. *Journal of Neuroscience*, *35*(49), 16064–16076. https://doi.org/10.1523/JNEUROSCI.2454-15.2015

Markowitz, D. A., Curtis, C. E., & Pesaran, B. (2015). Multiple component networks support working memory in prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(35), 11084–11089. https://doi.org/10.1073/pnas.1504172112

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X. J., & Freedman, D. J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature Neuroscience*, *22*(7), 1159–1167. https://doi.org/10.1038/s41593-019-0414-3

Mateeff, S., & Gourevich, A. (1983). Peripheral vision and perceived visual direction. *Biological Cybernetics*, *49*(2), 111–118. https://doi.org/10.1007/BF00320391

McAdams, C. J., & Maunsell, J. H. R. (2000). Attention to both space and feature modulates neuronal responses in macaque area V4. *Journal of Neurophysiology*, *83*(3), 1751–1755. https://doi.org/10.1152/jn.2000.83.3.1751

McKeown, D., & Mercer, T. (2012). Short-term forgetting without interference. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(4), 1057–1068. https://doi.org/10.1037/a0027749

McNab, F., & Dolan, R. J. (2014). Dissociating distractor-filtering at encoding and during maintenance. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 960–967. https://doi.org/10.1037/a0036013

McNab, F., & Klingberg, T. (2008). Prefrontal cortex and basal ganglia control access to working memory. *Nature Neuroscience*, *11*(1), 103–107. https://doi.org/10.1038/nn2024

McNab, F., Varrone, A., Farde, L., Jucaite, A., Bystritsky, P., Forssberg, H., & Klingberg, T. (2009). Changes in cortical dopamine D1 receptor binding associated with cognitive training. *Science*, *323*(5915), 800–802. https://doi.org/10.1126/science.1166102

Mendoza-Halliday, D., & Martinez-Trujillo, J. C. (2017). Neuronal population coding of perceived and memorized visual features in the lateral prefrontal cortex. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms15471

MercedesBenz. (2020). *How Does F1 Simulation Work?* https://www.mercedesamgf1.com/en/news/2020/portugal-grand-prix/how-does-f1-simulation-work/

Merchant, H., Fortes, A. F., & Georgopoulos, A. P. (2004). Short-term memory effects on the representation of two-dimensional space in the rhesus monkey. *Animal Cognition*, *7*(3), 133–143. https://doi.org/10.1007/s10071-003-0201-z

Mewhort, D. J. K., & Campbell, A. J. (1978). Processing spatial information and the selective-masking effect. *Perception & Psychophysics*, *24*(1), 93–101. https://doi.org/10.3758/BF03202978

Meyer, T., Qi, X. L., Stanford, T. R., & Constantinidis, C. (2011). Stimulus selectivity in dorsal and ventral prefrontal cortex after training in working memory tasks. *Journal of Neuroscience*, *31*(17), 6266–6276. https://doi.org/10.1523/JNEUROSCI.6798-10.2011

Meyers, E. M., Qi, X. L., & Constantinidis, C. (2012). Incorporation of new information into prefrontal cortical activity after learning working memory tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(12), 4651–4656. https://doi.org/10.1073/pnas.1201022109

Miller, E., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *16*(16), 5154–5167. https://doi.org/10.1523/JNEUROSCI.16-16-05154.1996

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Milner, B. (1963). Effects of Different Brain Lesions on Card Sorting: The Role of the Frontal Lobes. *Archives of Neurology*, *9*(1), 90–100. https://doi.org/10.1001/archneur.1963.00460070100010

Minami, T., & Inui, T. (2001). A neural network model of working memory processing of “what” and “where” information. *Lecture Notes in Computer Science*, *2084 LNCS*(PART 1), 126–133. https://doi.org/10.1007/3-540-45720-8_15

Mitrani, L., & Dimitrov, G. (1982). Retinal location and visual localization during pursuit eye movement. *Vision Research*, *22*(8), 1047–1051. https://doi.org/10.1016/0042-6989(82)90041-4

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic Theory of Working Memory. *Science*, *179*(March), 1543–1547.

Murray, J. D., Bernacchia, A., Roy, N. A., Constantinidis, C., Romo, R., & 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), 394–399. https://doi.org/10.1073/pnas.1619449114

Murray, J. D., Jaramillo, J., & Wang, X.-J. (2017). Working memory and decision-making in a frontoparietal circuit model. *Journal of Neuroscience*, *37*(50), 12167–12186. https://doi.org/10.1523/JNEUROSCI.0343-17.2017

Müsseler, J., Van Der Heijden, A. H. C., Mahmud, S. H., Deubel, H., & Ertsey, S. (1999). Relative mislocalization of briefly presented stimuli in the retinal periphery. *Perception and Psychophysics*, *61*(8), 1646–1661. https://doi.org/10.3758/BF03213124

N

Nakajima, M., Schmitt, L. I., & Halassa, M. M. (2019). Prefrontal Cortex Regulates Sensory Filtering through a Basal Ganglia-to-Thalamus Pathway. *Neuron*, *103*(3), 445-458.e10. https://doi.org/10.1016/j.neuron.2019.05.026

Nassar, M. R., Helmers, J. C., & Frank, M. J. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological Review*. https://doi.org/10.1037/rev0000101

O

O'Toole, B., & Wenderoth, P. (1977). The tilt illusion: Repulsion and attraction effects in the oblique meridian. *Vision Research*, *17*(3), 367–374. https://doi.org/10.1016/0042-6989(77)90025-6

Olesen, P. J., Westerberg, H., & Klingberg, T. (2004). Increased prefrontal and parietal activity after training of working memory. *Nature Neuroscience*, *7*(1), 75–79. https://doi.org/10.1038/nn1165

Opris, I., Barborica, A., & Ferrera, V. P. (2005). Effects of electrical microstimulation in monkey frontal eye field on saccades to remembered targets. *Vision Research*, *45*(27), 3414–3429. https://doi.org/10.1016/j.visres.2005.03.014

Osaka, N. (1977). Effect of refraction on perceived locus of a target in the peripheral visual field. *Journal of Psychology: Interdisciplinary and Applied*, *95*(1), 59–62. https://doi.org/10.1080/00223980.1977.9915860

P

Panichello, M. F., & Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature*, *592*(7855), 601–605. https://doi.org/10.1038/s41586-021-03390-w

Papadimitriou, C., Ferdoash, A., & Snyder, L. H. (2015). Ghosts in the machine: Memory interference from the previous trial. *Journal of Neurophysiology*, *113*(2), 567–577. https://doi.org/10.1152/jn.00402.2014

Park, B. Y., Byeon, K., & Park, H. (2019). FuNP (fusion of neuroimaging preprocessing) pipelines: A fully automated preprocessing software for functional magnetic resonance imaging. *Frontiers in Neuroinformatics*, *13*, 5. https://doi.org/10.3389/fninf.2019.00005

Parthasarathy, A., Herikstad, R., Bong, J. H., Medina, F. S., Libedinsky, C., & Yen, S. C. (2017). Mixed selectivity morphs population codes in prefrontal cortex. *Nature Neuroscience*. https://doi.org/10.1038/s41593-017-0003-2

Pasternak, T., & Zaksas, D. (2003). Stimulus specificity and temporal dynamics of working memory for visual motion. *Journal of Neurophysiology*, *90*(4), 2757–2762. https://doi.org/10.1152/jn.00422.2003

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pertzov, Y., Bays, P. M., Joseph, S., & Husain, M. (2013). Rapid forgetting prevented by retrospective attention cues. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(5), 1224–1231. https://doi.org/10.1037/a0030947

Pesaran, B., Pezaris, J. S., Sahani, M., Mitra, P. P., & Andersen, R. A. (2002). Temporal structure in neuronal activity during working memory in macaque parietal cortex. *Nature Neuroscience*, *5*(8), 805–811. https://doi.org/10.1038/nn890

Ploner, C. J., Gaymard, B., Rivaud, S., Agid, Y., & Pierrot-Deseilligny, C. (1998). Temporal limits of spatial working memory in humans. *European Journal of Neuroscience*, *10*(2), 794–797. https://doi.org/10.1046/j.1460-9568.1998.00101.x

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology Quarterly Journal of EJrperimental Psychology*, *32*(32), 3–25. https://doi.org/10.1080/00335558008248231

Posner, M. I., & Gilbert, C. D. (1999). Attention and primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(6), 2585–2587. https://doi.org/10.1073/pnas.96.6.2585

Pouget, A., Deneve, S., Ducom, J. C., & Latham, P. E. (1999). Narrow versus wide tuning curves: What's best for a population code? *Neural Computation*, *11*(1), 85–90. https://doi.org/10.1162/089976699300016818

Pratte, M. S., Park, Y. E., Rademaker, R. L., & Tong, F. (2017). Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 6–17. https://doi.org/10.1037/xhp0000302

Pribram, K. H., Ahumada, A., Hartog, J., & Ross, L. (1964). A progress report on the neurological process disturbed by frontal lesions in primates. In W. I. M. & K. Akert (Eds.), *The frontal granular cortex and behavior.* (pp. 28–55). McGraw-Hill Rook.

Pribram, Karl H., Mishkin, M., Rosvold, H. E., & Kaplan, S. J. (1952). Effects on delayed-response performance of lesions of dorsolateral and ventromedial frontal cortex of baboons. *Journal of Comparative and Physiological Psychology*, *45*(6), 565–575. https://doi.org/10.1037/h0061240

Q

Qi, X. L., & Constantinidis, C. (2012a). Correlated discharges in the primate prefrontal cortex before and after working memory training. *European Journal of Neuroscience*, *36*(11), 3538–3548. https://doi.org/10.1111/j.1460-9568.2012.08267.x

Qi, X. L., & Constantinidis, C. (2012b). Variability of prefrontal neuronal discharges before and after training in a working memory task. *PLoS ONE*, *7*(7), 1–8. https://doi.org/10.1371/journal.pone.0041053

Qi, X. L., Liu, R., Singh, B., Bestue, D., Compte, A., Vazdarjanova, A. I., Blake, D. T., & Constantinidis, C. (2021). Nucleus basalis stimulation enhances working memory by stabilizing stimulus representations in primate prefrontal cortical activity. *Cell Reports*, *36*(5). https://doi.org/10.1016/j.celrep.2021.109469

R

Rademaker, R. L., Bloem, I. M., De Weerd, P., & 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), 1650–1665. https://doi.org/10.1037/xhp0000110

Rademaker, R. L., Chunharas, C., & Serences, J. T. (2019). Coexisting representations of sensory and mnemonic information in human visual cortex. *Nature Neuroscience*, *22*(8), 1336–1344. https://doi.org/10.1038/s41593-019-0428-x

Rahman, F., Nanu, R., Schneider, N. A., Katz, D., Lisman, J., & Pi, H. J. (2021). Optogenetic perturbation of projections from thalamic nucleus reuniens to hippocampus disrupts spatial working memory retrieval more than encoding. *Neurobiology of Learning and Memory*, *179*. https://doi.org/10.1016/j.nlm.2021.107396

Raiguel, S., Vogels, R., Mysore, S. G., & Orban, G. A. (2006). Learning to see the difference specifically alters the most informative V4 neurons. *Journal of Neuroscience*, *26*(24), 6589–6602. https://doi.org/10.1523/JNEUROSCI.0457-06.2006

Rainer, G., Asaad, W. F., & Miller, E. (1998). Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature*, *393*(6685), 577–579. https://doi.org/10.1038/31235

Ramaraju, S., Roula, M. A., & McCarthy, P. W. (2020). Transcranial direct current stimulation and working memory: Comparison of effect on learning shapes and English letters. *PLoS ONE*, *15*(7 July). https://doi.org/10.1371/journal.pone.0222688

Reynolds, J. H., & Chelazzi, L. (2004). Attentional modulation of visual processing. In *Annual Review of Neuroscience* (Vol. 27, pp. 611–647). Annu Rev Neurosci. https://doi.org/10.1146/annurev.neuro.26.041002.131039

Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, *26*(3), 703–714. https://doi.org/10.1016/S0896-6273(00)81206-4

Rose, N. S., Larocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316), 1136–1140.

Ross, J., Morrone, M. C., & Burr, D. C. (1997). Compression of visual space before saccades. *Nature*, *386*(6625), 598–601. https://doi.org/10.1038/386598a0

Rovamo, J., & Virsu, V. (1979). An estimation and application of the human cortical magnification factor. *Experimental Brain Research*, *37*(3), 495–510. https://doi.org/10.1007/BF00236819

Rovamo, Jyrki, Virsu, V., & Näsänen, R. (1978). Cortical magnification factor predicts the photopic contrast sensitivity of peripheral vision. *Nature*, *271*(5640), 54–56. https://doi.org/10.1038/271054a0

S

Sakai, K., Rowe, J. B., & Passingham, R. E. (2002). Active maintenance in prefrontal area 46 creates distractor-resistant memory. *Nature Neuroscience*, *5*(5), 479–484. https://doi.org/10.1038/nn846

Salthouse, T. A. (2014). Relations between running memory and fluid intelligence. *Intelligence*, *43*(1), 1–7. https://doi.org/10.1016/j.intell.2013.12.002

Sanayei, M., Chen, X., Chicharro, D., Distler, C., Panzeri, S., & Thiele, A. (2018). Perceptual learning of fine contrast discrimination changes neuronal tuning and population coding in macaque V4. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-06698-w

Sandberg, A., Tegnér, J., & Lansner, A. (2003). A working memory model based on fast Hebbian learning. *Network: Computation in Neural Systems*, *14*(4), 789–802. https://doi.org/10.1088/0954-898X_14_4_309

Sattari, N., Whitehurst, L. N., Ahmadi, M., & Mednick, S. C. (2019). Does working memory improvement benefit from sleep in older adults? *Neurobiology of Sleep and Circadian Rhythms*, *6*(December 2018), 53–61. https://doi.org/10.1016/j.nbscr.2019.01.001

Schutte, A. R., & DeGirolamo, G. J. (2020). Test of a dynamic neural field model: spatial working memory is biased away from distractors. *Psychological Research*, *84*(6), 1528–1544. https://doi.org/10.1007/s00426-019-01166-6

Scimeca, J. M., Kiyonaga, A., & D'Esposito, M. (2018). Reaffirming the Sensory Recruitment Account of Working Memory. In *Trends in Cognitive Sciences* (Vol. 22, Issue 3, pp. 190–192). Trends Cogn Sci. https://doi.org/10.1016/j.tics.2017.12.007

Seeholzer, A., Deger, M., & Gerstner, W. (2019). Stability of working memory in continuous attractor networks under the control of shortterm plasticity. *PLoS Computational Biology*, *15*(4). https://doi.org/10.1371/journal.pcbi.1006928

Serences, J. T. (2016). Neural mechanisms of information storage in visual short-term memory. *Vision Research*, *128*, 53–67. https://doi.org/10.1016/j.visres.2016.09.010

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science*, *20*(2), 207–214. https://doi.org/10.1111/j.1467-9280.2009.02276.x

Seung, H. S. (2000). Half a century of Hebb. *Nature Neuroscience*, *3*(11s), 1166. https://doi.org/10.1038/81430

Sheth, B. R., & Shimojo, S. (2001). Compression of space in visual memory. *Vision Research*, *41*(3), 329–341. https://doi.org/10.1016/S0042-6989(00)00230-3

Shin, H., Zou, Q., & Ma, W. J. (2017). The effects of delay duration on visual working memory for orientation. *Journal of Vision*, *17*(14), 1–24. https://doi.org/10.1167/17.14.10

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working Memory Capacity and Fluid Intelligence: Maintenance and Disengagement. *Perspectives on Psychological Science*, *11*(6), 771–799. https://doi.org/10.1177/1745691616650647

Spaak, E., Watanabe, K., Funahashi, S., & Stokes, M. G. (2017). Stable and dynamic coding for working memory in primate prefrontal cortex. *Journal of Neuroscience*, *37*(27), 6503–6516. https://doi.org/10.1523/JNEUROSCI.3364-16.2017

Spencer, J. P., & Hund, A. M. (2002). Prototypes and particulars: Geometric and experience-dependent spatial categories. *Journal of Experimental Psychology: General*, *131*(1), 16–37. https://doi.org/10.1037/0096-3445.131.1.16

Spinelli, S., Ballard, T., Feldon, J., Higgins, G. A., & Pryce, C. R. (2006). Enhancing effects of nicotine and impairing effects of scopolamine on distinct aspects of performance in computerized attention and working memory tasks in marmoset monkeys. *Neuropharmacology*, *51*(2), 238–250. https://doi.org/10.1016/j.neuropharm.2006.03.012

Spinks, J. A., Zhang, J. X., Fox, P. T., Gao, J. H., & Hai Tan, L. (2004). More workload on the central executive of working memory, less attention capture by novel visual distractors: Evidence from an fMRI study. *NeuroImage*, *23*(2), 517–524. https://doi.org/10.1016/j.neuroimage.2004.06.025

Sprague, T. C., Ester, E. F., & Serences, J. T. (2014). Reconstructions of information in visual spatial working memory degrade with memory load. *Current Biology*, *24*(18), 2174–2180. https://doi.org/10.1016/j.cub.2014.07.066

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring Latent Visual Working Memory Representations in Human Cortex. *Neuron*, *91*(3), 694–707. https://doi.org/10.1016/j.neuron.2016.07.006

Staugaard, C. F., Petersen, A., & Vangkilde, S. (2016). Eccentricity effects in vision and attention. *Neuropsychologia*, *92*, 69–78. https://doi.org/10.1016/j.neuropsychologia.2016.06.020

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. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-18033-3

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, *78*(2), 364–375. https://doi.org/10.1016/j.neuron.2013.01.039

Sugase-Miyamoto, Y., Liu, Z., Wiener, M. C., Optican, L. M., & Richmond, B. J. (2008). Short-term memory trace in rapidly adapting synapses of inferior temporal cortex. *PLoS Computational Biology*, *4*(5). https://doi.org/10.1371/journal.pcbi.1000073

Sun, Y., Yang, Y., Galvin, V. C., Yang, S., Arnsten, A. F., & Wang, M. (2017). Nicotinic α4β2 cholinergic receptor influences on dorsolateral prefrontal cortical neuronal firing during a working memory task. *Journal of Neuroscience*, *37*(21), 5366–5377. https://doi.org/10.1523/JNEUROSCI.0364-17.2017

Supèr, H., Spekreijse, H., & Lamme, V. A. F. (2001). A neural correlate of working memory in the monkey primary visual cortex. *Science*, *293*(5527), 120–124. https://doi.org/10.1126/science.1060496

Suzuki, M., & Gottlieb, J. (2013). Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nature Neuroscience*, *16*(1), 98–104. https://doi.org/10.1038/nn.3282

T

Takeuchi, T., Duszkiewicz, A. J., Sonneborn, A., Spooner, P. A., Yamasaki, M., Watanabe, M., Smith, C. C., Fernández, G., Deisseroth, K., Greene, R. W., & Morris, R. G. M. (2016). Locus coeruleus and dopaminergic consolidation of everyday memory. *Nature*. https://doi.org/10.1038/nature19325

Tang, H., Qi, X. L., Riley, M. R., & Constantinidis, C. (2019). Working memory capacity is enhanced by distributed prefrontal activation and invariant temporal dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(14), 7095–7100. https://doi.org/10.1073/pnas.1817278116

Teng, C., & Kravitz, D. J. (2019). Visual working memory directly alters perception. In *Nature Human Behaviour* (Vol. 3, Issue 8, pp. 827–836). Nat Hum Behav. https://doi.org/10.1038/s41562-019-0640-4

Toet, A., & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research*, *32*(7), 1349–1357. https://doi.org/10.1016/0042-6989(92)90227-A

Townsend, V. M. (1973). Loss of spatial and identity information following a tachistoscopic exposure. *Journal of Experimental Psychology*, *98*(1), 113–118. https://doi.org/10.1037/h0034309

Treue, S., & Maunsell, J. H. R. (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature*, *382*(6591), 539–541. https://doi.org/10.1038/382539a0

Tsodyks, M., Pawelzik, K., & Markram, H. (1998). Neural Networks with Dynamic Synapses. *Neural Computation*, *10*(4), 821–835. https://doi.org/10.1162/089976698300017502

Tsujimoto, S., & Sawaguchi, T. (2004). Properties of delay-period neuronal activity in the primate prefrontal cortex during memory- and sensory-guided saccade tasks. *European Journal of Neuroscience*, *19*(2), 447–457. https://doi.org/10.1111/j.0953-816X.2003.03130.x

U

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, *64*(1), 49–60. https://doi.org/10.1037/h0044616

V

Vaidya, A. R., Pujara, M. S., Petrides, M., Murray, E. A., & Fellows, L. K. (2019). Lesion Studies in Contemporary Neuroscience. In *Trends in Cognitive Sciences* (Vol. 23, Issue 8, pp. 653–671). Trends Cogn Sci. https://doi.org/10.1016/j.tics.2019.05.009

Van Der Stigchel, S., Merten, H., Meeter, M., & Theeuwes, J. (2007). The effects of a task-irrelevant visual event on spatial working memory. *Psychonomic Bulletin and Review*, *14*(6), 1066–1071. https://doi.org/10.3758/BF03193092

Van Ede, F., Chekroud, S. R., Stokes, M. G., & Nobre, A. C. (2018). Decoding the influence of anticipatory states on visual perception in the presence of temporal distractors. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-03960-z

Van Essen, D. C., Newsome, W. T., & Maunsell, J. H. R. (1984). The visual field representation in striate cortex of the macaque monkey: Asymmetries, anisotropies, and individual variability. *Vision Research*, *24*(5), 429–448. https://doi.org/10.1016/0042-6989(84)90041-5

Vergauwe, E., Barrouillet, P., & Camos, V. (2010). Do mental processes share a domain-general resource? *Psychological Science*, *21*(3), 384–390. https://doi.org/10.1177/0956797610361340

Vijayraghavan, S., Major, A. J., & Everling, S. (2017). Neuromodulation of prefrontal cortex in non-human primates by dopaminergic receptors during rule-guided flexible behavior and cognitive control. *Frontiers in Neural Circuits*, *11*. https://doi.org/10.3389/fncir.2017.00091

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, *438*(7067), 500–503. https://doi.org/10.1038/nature04171

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(6), 1436–1451. https://doi.org/10.1037/0096-1523.32.6.1436

Voytek, B., & Knight, R. T. (2010). Prefrontal cortex and basal ganglia contributions to visual working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(42), 18167–18172. https://doi.org/10.1073/pnas.1007277107

W

Wade, N. J., Spillmann, L., & Swanston, M. T. (1996). Visual motion aftereffects: Critical adaptation and test conditions. *Vision Research*, *36*(14), 2167–2175. https://doi.org/10.1016/0042-6989(95)00266-9

Wan, Q., Cai, Y., Samaha, J., & Postle, B. R. (2020). Tracking stimulus representation across a 2-back visual working memory task. *Royal Society Open Science*, *7*(8), 190228. https://doi.org/10.1098/rsos.190228

Wang, K. W., Ye, X. L., Huang, T., Yang, X. F., & Zou, L. Y. (2019). Optogenetics-induced activation of glutamate receptors improves memory function in mice with Alzheimer's disease. *Neural Regeneration Research*, *14*(12), 2147–2155. https://doi.org/10.4103/1673-5374.262593

Wang, X.-J., Tegnér, J., Constantinidis, C., & Goldman-Rakic, P. S. (2004). Division of labor among distinct subtypes of inhibitory neurons in a cortical microcircuit of working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(5), 1368–1373. https://doi.org/10.1073/pnas.0305337101

Warren, J. M., Leary, R. W., Harlow, H. F., & French, G. M. (1957). Function of association cortex in monkeys. *The British Journal of Animal Behaviour*, *5*(4), 131–138. https://doi.org/10.1016/S0950-5601(57)80019-7

Watanabe, K., & Funahashi, S. (2014). Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. *Nature Neuroscience*, *17*(4), 601–611. https://doi.org/10.1038/nn.3667

Wei, X. X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. *Nature Neuroscience*, *18*(10), 1509–1517. https://doi.org/10.1038/nn.4105

Wei, Z., Wang, X.-J., & Wang, D. H. (2012). From distributed resources to limited slots in multiple-item working memory: A spiking network model with normalization. *Journal of Neuroscience*, *32*(33), 11228–11240. https://doi.org/10.1523/JNEUROSCI.0735-12.2012

White, J. M., Sparks, D. L., & Stanford, T. R. (1994). Saccades to remembered target locations: an analysis of systematic and variable errors. *Vision Research*, *34*(1), 79–92. https://doi.org/10.1016/0042-6989(94)90259-3

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 1120–1135. https://doi.org/10.1167/4.12.11

Wilson, H. R., & Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, *13*(2), 55–80. https://doi.org/10.1007/BF00288786

Wimmer, K., Nykamp, D. Q., Constantinidis, C., & Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, *17*(3), 431–439. https://doi.org/10.1038/nn.3645

Wolf, C., & Lappe, M. (2021). Salient objects dominate the central fixation bias when orienting toward images. *Journal of Vision*, *21*(8), 1–21. https://doi.org/10.1167/jov.21.8.23

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, *20*(6), 864–871. https://doi.org/10.1038/nn.4546

Worsley, K. J., & Friston, K. J. (1995). Analysis of fMRI time-series revisited — Again. *NeuroImage*, *2*(3), 173–181. https://doi.org/10.1006/nimg.1995.1023

X

Xu, Y. (2018). Sensory Cortex Is Nonessential in Working Memory Storage. In *Trends in Cognitive Sciences* (Vol. 22, Issue 3, pp. 192–193). Elsevier Ltd. https://doi.org/10.1016/j.tics.2017.12.008

Y

Yang, T., & Maunsell, J. H. R. (2004). The Effect of Perceptual Learning on Neuronal Responses in Monkey Visual Area V4. *Journal of Neuroscience*, *24*(7), 1617–1626. https://doi.org/10.1523/JNEUROSCI.4442-03.2004

Yang, Y., Paspalas, C. D., Jin, L. E., Picciotto, M. R., Arnsten, A. F. T., & Wang, M. (2013). Nicotinic α7 receptors enhance NMDA cognitive circuits in dorsolateral prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(29), 12078–12083. https://doi.org/10.1073/pnas.1307849110

Yoon, J. H., Curtis, C. E., & D'Esposito, M. (2006). Differential effects of distraction during working memory on delay-period activity in the prefrontal cortex and the visual association cortex. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2005.08.024

Yue, J., Zhong, S., Luo, A., Lai, S., He, T., Luo, Y., Wang, Y., Zhang, Y., Shen, S., Huang, H., Wen, S., & Jia, Y. (2021). Correlations between working memory impairment and neurometabolites of the prefrontal cortex in drug-naive obsessive-compulsive disorder. *Neuropsychiatric Disease and Treatment*, *17*, 2647–2657. https://doi.org/10.2147/NDT.S296488

Z

Zelinski, E. L. (2016). *THE INFLUENCE OF CONTEXT AND STRATEGY ON SPATIAL TASK PERFORMANCE*. University of Lethbridge.

Zhang, K., & Sejnowski, T. J. (1999). Neuronal tuning: To sharpen or broaden? *Neural Computation*, *11*(1), 75–84. https://doi.org/10.1162/089976699300016809

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. https://doi.org/10.1038/nature06860

Zhang, Y., Cao, L., Varga, V., Jing, M., Karadas, M., Li, Y., & Buzsáki, G. (2021). Cholinergic suppression of hippocampal sharp-wave ripples impairs working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(15). https://doi.org/10.1073/pnas.2016432118

Abbreviations

List of abbreviations

Abbreviation	Meaning
Ach	acetylcholine
BOLD	blood oxygenation level lependent
ccw	counterclockwise
ci	confidence interval
cw	clockwise
dlPFC	dorsolateral prefrontal cortex
EPI	echo planar imaging
fMRI	functional magnetic resonance imaging
FWHM	full width at half maximum
gF	general fluid intelligence
IEM	Inverted encoding model
IQ	intelligence quotient
ITI	inter-trial interval
IPS	Intraparietal sulcus
LIP	lateral intraparietal
MNI	Montreal Neurological Institute
MRI	magnetic resonance imaging
NB	Nucleus Basalis of Meynert
NT	non-target
ODR	oculomotor delayed-response
PA	persistent activity
PFC	prefrontal cortx
RF	receptive field
ROI	region of interest
sem	standard error of the mean
SOA	stimulus onset asynchrony
SPM	Statistical Parametric Mapping
sPCS	Superior precentral sulcus
SRT	sensory recruitment theory
STD	short-term synaptic depression

STF	short-term synaptic facilitation
STP	short-term synaptic plasticity
TDOA	target-distractor onset asynchrony
TE	time to echo
TR	repetition time
vsWM	visuospatial working memory
WM	working memory

Neuron playing chess

@WOMBO by Ester Barrios